Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

	04/24/2025
Zhuohan Chi	Date

Early Detection of Neonatal Infection in NICU Using Machine Learning Models

By

Zhuohan Chi

MPH

Rollins School of Public Health, Emory University

Department of Biostatistics and Bioinformatics

Zhaohui Qin, PhD

Committee Chair

Steve Pittard, PhD

Committee Member

Early Detection of Neonatal Infection in NICU Using Machine Learning Models

By

Zhuohan Chi

Bachelor of Arts, Mathematics,

Boston University, 2022

Committee Chair: Zhaohui Qin, PhD

An abstract of a thesis submitted to the Faculty of the Rollins School of Public Health of Emory University in partial fulfillment of the requirements for the degree of Master of Public Health in Department of Biostatistics and Bioinformatics

2025

Abstract

Early Detection of Neonatal Infection in NICU Using Machine Learning Models

By

Zhuohan Chi

Early and accurate detection of neonatal infection is critical but current scores (SNAP-II, CRIB) require 12–24 h of data. Using the MIMIC-III NICU cohort, we built an explainable ML pipeline that aggregates vital-sign and lab data from the first 30 min and 120 min after admission. Missing values were handled with iterative multivariate imputation, and 18 classifiers were tuned via stratified 5-fold cross-validation. CatBoost performed best at 30 min (F1 = 0.76; Acc = 0.79), while Gradient Boosting led at 120 min (F1 = 0.80; Acc = 0.81), both surpassing traditional scores by ~6 pp. Feature-coverage experiments showed that half of the top features (30 min) and 80 % (120 min) maintained peak accuracy, yielding compact, interpretable models. A two-stage deployment—CatBoost for ultra-early screening, Gradient Boosting for refinement—could provide clinicians with reliable infection-risk alerts hours before conventional methods, supporting faster intervention and improved neonatal outcomes.

Early Detection of Neonatal Infection in NICU Using Machine Learning Models

By

Zhuohan Chi

Bachelor of Arts, Mathematics

Boston University, 2022

Committee Chair: Zhaohui Qin, PhD

A thesis submitted to the Faculty of the Rollins School of Public Health of Emory University in partial fulfillment of the requirements for the degree of Master of Public Health in Department of Biostatistics and Bioinformatics

2025

Table of Contents

CHAPTER 1: INTRODUCTION

Neonatal infections, including sepsis, remain a leading cause of morbidity and mortality in newborns, particularly in critical care settings such as neonatal intensive care units (NICUs)[1],[2]. These infections, often difficult to diagnose in their early stages, require timely identification to enable effective interventions. Delayed or inaccurate diagnosis can result in rapid disease progression, increased mortality, and long-term complications[3]. In parallel, antimicrobial resistance (AMR) has further complicated neonatal infection management, necessitating data-driven approaches to optimize diagnosis and treatment strategies[4].

Traditional scoring systems, such as SNAP-II and CRIB, are widely used to evaluate neonatal illness severity and predict outcomes[5],[6]. However, these tools are limited by their reliance on fixed variables and retrospective observations, often spanning 12–24 hours, which delays critical decision-making[7]. Furthermore, they lack the capacity to handle complex, multidimensional datasets or account for dynamic changes in patient conditions[8]. These limitations highlight the need for advanced, real-time predictive tools.

Machine learning (ML) has emerged as a transformative approach in healthcare, offering the ability to integrate and analyze complex data from diverse sources[9]. Leveraging ML and deep learning (DL) models, such as random forests, support vector machines, and recurrent neural networks, enables accurate predictions of neonatal

infection onset and associated mortality risks[10],[11]. These models have demonstrated superior performance compared to traditional methods, particularly when applied to large datasets such as MIMIC-III, a publicly available database of critical care patient records[12]. MIMIC-III provides a rich repository of neonatal data, including physiological measurements, laboratory results, and clinical interventions, making it an ideal resource for developing robust predictive models[13].

This thesis aims to develop a machine learning framework using data from MIMIC-III to predict neonatal infections and their associated mortality risks. By focusing on real-time measurement data collected within the critical first hours of care, the proposed approach seeks to enable early, actionable predictions. This framework will incorporate explainable AI methodologies to ensure transparency and trust in the predictive models, facilitating their integration into clinical workflows[14]. The outcomes of this research have the potential to enhance early diagnosis, improve resource allocation, and ultimately reduce neonatal mortality in NICUs.

CHAPTER 2: METHOD

Figure 2 introduced the research pipeline. It begins by selecting NICU patients who were admitted only once to the ICU and survived at least 120 minutes post-admission. Relevant clinical data are extracted from chartevents, labevents, and selected variables, then aggregated into 30- and 120-minute windows. After cleaning and imputing missing values, machine learning models are developed through cross-validation, with hyperparameters optimized via grid search. Finally, feature selection and SHAP analysis are used to interpret model predictions and highlight important clinical variables.

2.1 Data Source

The data source of this study is Medical Information Mart for Intensive Care III (MIMIC-III), a relational critical care database developed by the MIT Lab for Computational Physiology. MIMIC-III is a publicly available dataset containing de-identified health data from patients admitted to critical care units at Beth Israel Deaconess Medical Center. It contains detailed health-related data for 46,520 unique critical care patients admitted between 2001 and 2013, covering 58,976 hospital admissions and 38597 ICU stays. The dataset includes demographics, clinical diagnoses, procedures, medications, lab results, etc. MIMIC-III contains over 2 million chart events, 380,000 laboratory measurements, and over 10,000 procedures, making it a comprehensive dataset for critical care research.

2.2 Data Preprocessing

After data collection, preprocessing is crucial for achieving reliable, complete, and perfect data for prediction tasks[15],[16]. In this research, preprocessing includes gathering relative data and dealing with missing values.

This research aims to create a machine learning model to predict newborn infections within a short period of time. The first step is gathering patients in the NICU, which are the natal patients. We exclude patients been admitted to ICU more than once, and patients dead in two hours after being admitted to ICU for reducing the complexity of model[17].

Various physiological and hematological parameters are essential for early identification and management in predicting neonatal infections. Heart rate is a key indicator; tachycardia may signal systemic infections or sepsis, while bradycardia can indicate severe infection. Respiratory rate is also crucial, with tachypnea reflecting respiratory distress and decreased rates indicating failure[18]. Oxygen saturation (SaO_2) levels reveal respiratory efficiency, particularly during infections such as pneumonia[19]. Temperature regulation is vital, as fever or hypothermia indicates systemic responses to infection. Elevated temperatures suggest inflammation and hypothermia often indicates severe sepsis. Blood pressure readings are important for assessing[20] cardiovascular stability, with hypotension being a critical late sign of septic shock[21]. Hematological markers are indispensable; white blood cell counts can indicate infection, while neutrophil counts reflect acute bacterial responses. Lymphocyte

counts may decrease, signifying immunosuppression. Thrombocytopenia and abnormal hemoglobin levels may complicate infection management[22]. Immature granulocytes in blood smears indicate a strong bone marrow response to severe infections[23].

All measurements are not always collected simultaneously for each patient, and measurements may be collected multiple times, we aggregate the data for one patient by two time windows, 30 minutes and 120 minutes. In the 30 minutes dataset, we aggregate the data to minimum, maximum, and mean for each variable in first 30 minutes after patients admitted to ICU. Then do the same thing for the 120 minutes dataset.

2.3 Missing Value Imputation

In a hospital environment and specifically in ICUs, many variables are measured. Figure 2 shows the missing value distribution in the datasets. However, these measures are not always conducted and available at the same time for any patient and this is the reason for the frequent problem of missing data, especially in early hours of admission. Appropriate handling of missing data is crucial to ensure the validity and generalizability of study findings[24],[25].

There are three missing types: Missing Completely at Random, missing at random, and missing not at random. Understanding the types of missing data is necessary to determine whether imputation methods, sensitivity analyses, or model adjustments are necessary to avoid misleading conclusions in research findings[26].

We explored the nature of missingness through a multi-step approach. First, we employed missingno visualizations—a matrix plot to reveal patterns of missing data in observations and variables, and a heatmap to indicate correlations in missingness[27]. Next, we performed Little's MCAR test by label encoding categorical variables and applying mean imputation to derive a complete covariance matrix. We compared this covariance matrix to that of the fully observed subset and used the resulting statistics to ascertain whether the data were missing completely at random (MCAR). A non-significant p-value (above 0.05) supported the MCAR hypothesis, while a significant result indicated that the data might follow missing at random (MAR) or missing not at random (MNAR) mechanisms[28].

To differentiate between MAR and MNAR, we conducted chi-square tests for each variable with missing data, creating a binary indicator to evaluate associations with other observed variables; significant results suggested MAR[29]. Finally, we performed two-sample t-tests to determine if observed values differed between the "missing" and "non-missing" groups for each variable. A significant difference ($p < 0.05$) suggested that missingness could be MNAR, implying dependence on unobserved information rather than solely on observed data.

We examined two strategies for addressing missing data—row removal based on a pre-defined missing threshold (θ) and three imputation methods (Simple Imputer, KNN Imputer, and Iterative Imputer)[30],[31]. We systematically varied θ from 0.95 to 0.6 in increments of 0.05, discarding rows that surpassed each threshold's proportion of

missing values to ensure that highly incomplete rows did not compromise imputation results. Subsequently, we artificially masked 10% of the numeric entries to create a known “ground truth” for performance evaluation, comparing original values to the imputed ones.

We assessed imputation quality using mean squared error (MSE) and mean absolute error (MAE), calculated over all features where masking occurred. Hyperparameters like k-values for KNN and iteration counts for the Iterative Imputer were tuned via grid search, providing a robust benchmark of each method’s effectiveness. By systematically quantifying the trade-offs between excluding rows and employing various imputation techniques, this approach provides a clearer understanding of optimal strategies for handling incomplete datasets.

2.4 Model Selection

We explored a diverse range of classification algorithms to capture a broad spectrum of decision boundaries, incorporating models from multiple learning paradigms, such as Linear and Statistical Models (Logistic Regression, Naïve Bayes), Tree-Based Methods (Decision Tree, Random Forest, Gradient Boosting, AdaBoost, XGBoost, LightGBM, Extra Trees, CatBoost, HistGradientBoosting, and Bagging Classifier), Instance-Based Learning (k-Nearest Neighbors), Neural Networks (Multilayer Perceptron), Support Vector Machine[32].

This comprehensive set of models ensured that both linear and non-linear decision boundaries were considered, allowing for a more robust evaluation of classification performance.

To ensure a reliable assessment of model generalizability, we adopted a Stratified K-Fold cross-validation approach with five folds (n_splits = 5). This technique maintains class distribution consistency across folds, mitigating potential class imbalance issues and ensuring fair evaluation across all candidate models[33].

Each model's predictive performance was assessed using four key metrics, computed through cross-validation:

- Accuracy: The proportion of correctly classified instances.
- Precision (weighted): The class-weighted mean of precision across all categories.
- Recall (weighted): The class-weighted means of recall, ensuring balanced sensitivity.
- F1-Score (weighted): The harmonic means of precision and recall, providing a robust measure of model effectiveness.

The mean scores for each metric were computed across all five folds, ensuring stability in performance evaluation.

The best-performing model was determined based on its mean F1-score, as this metric provides a balanced evaluation of both precision and recall, making it particularly suitable for imbalanced datasets[34]. Once the top-performing model was identified, it was re-trained on the entire imputed dataset (X_imputed, y) to produce a final, optimized predictive model ready for validation and deployment.

2.5 Hyperparameter Optimization

We rigorously evaluated a diverse set of algorithms by employing this multi-step framework—comprising imputation, cross-validation, and metric-based model selection. This approach ensured that the final chosen model exhibited strong predictive performance across multiple evaluation dimensions, enhancing its reliability for real-world applications[35].

A predefined grid of hyperparameter values is established, covering a range of model-specific configurations such as learning rate, regularization strength, and complexity parameters. The grid systematically enumerates all possible combinations to be tested, ensuring that various configurations are explored comprehensively.

For evaluation of each hyperparameter combination, a k-fold cross-validation procedure was employed. Specifically, the training data were split into k folds (e.g., k=5), and each fold in turn was treated as a temporary validation set while the remaining folds were used for model training. Five folds offered the best trade-off between statistical reliability and computation cost. Performance metrics, such as accuracy, precision, recall, or F1-score, were computed on the validation fold, and the process repeated for each combination of parameters. This approach mitigated the risk of overfitting to a single train–validation partition and provided more robust estimates of out-of-sample performance[36].

All hyperparameter combinations in the predefined grid were iterated over, with each model built and evaluated using the cross-validation scheme described above. For

Gradient Boosting, the grid spanned n_estimators, learning_rate, max_depth, subsample, and min_samples_leaf to balance ensemble size, step size, tree complexity, and sampling noise. For CatBoost, we varied iterations, learning_rate, depth, l2_leaf_reg, and bagging_temperature, allowing fine-grained control over boosting rounds, tree depth, regularization strength, and stochasticity. The mean (and optionally standard deviation) of the performance metrics across folds was recorded for each parameter setting. Computational parallelization (using multiple CPU cores or threads) was leveraged where it was feasible to expedite the search process. The configuration achieving the highest mean performance metric (e.g., the highest mean F1-score) was identified as optimal. In certain cases, a secondary performance measure (e.g., model complexity, run time) or a tiebreaker (e.g., validation loss) was used to distinguish between comparably performing solutions.

2.6 Model Selection

The selected model's built-in feature importance mechanism was first used to assign each variable a numerical score reflecting its contribution to predictive accuracy. The full feature matrix was used to train baseline tree-based ensembles, including Random Forest, Extra Trees, Gradient Boosting, HistGradient Boosting, LightGBM, XGBoost, and CatBoost. The native importance metric of each model—mean decrease in impurity for bagging trees, total gain for boosting algorithms—was then extracted. These scores were min-max scaled to 0–1 and averaged to create a consensus ranking that highlighted variables deemed influential across models. We

then generated a ranked list of variables in descending order of importance. Subsequently, we employed the SHAP (SHapley Additive exPlanations) framework to quantify each feature's contribution to individual predictions, producing both summary plots and class-level insights. This interpretability analysis helped clarify which variables most strongly influenced the model's predictive outcomes.

To further investigate the relationship between feature subset size and model performance, we adopted an incremental coverage scheme ranging from 10% to 80% (in 5% increments). Specifically, if there were F total features, a coverage level of c% dictated retaining the top [c%×F] features according to the importance ranking. For example, a coverage of 10% preserved only the top 10% of features, whereas 80% coverage retained 80% of the most important ones. At each coverage level, we trained a new CatBoost model (using the same hyperparameters) and conducted a 5-fold stratified cross-validation. The mean and standard deviation of accuracy for each subset provided insight into both performance and stability across different feature inclusion thresholds.

CHAPTER 3: RESULT

Table 1 shows the final list of 23 selected variables. In the 30-minute dataset, several variables exhibit notably high missingness rates (e.g., Atypical Lymphocytes, Bands, Basophils, and other hematological markers often missing in over 70% of cases), while respiratory- and temperature-related features (e.g., Heart Rate, Resp Rate, Temp Skin [C]) show comparatively lower missingness (40–70%). Overall, most mean values differ between infected and non-infected groups in the 30-minute window, with parameters like heart rate, neutrophils, and white blood cells tending to show higher average values in the infected cohort. By contrast, the 120-minute dataset contains fewer missing observations across many features (with most falling in the 20–50% range) and maintains a similar pattern of differences between infected and non-infected groups (e.g., higher Heart Rate, lower Neutrophils among infected). Although the mean values of several variables (e.g., Basophils, Monocytes) remain comparable across groups, both time windows demonstrate consistent distinctions in key physiological and hematologic parameters between infected and non-infected neonates.

After doing Little's MCAR test, the p-value for both a 30-minute dataset and 120-minute dataset is less than 0.001, which indicates the missing data is not completely random. Then we utilize random forest accuracy to each variable, the result shows every row have a high accuracy (mean = 0.99).

3.1 Row Removal Threshold Result

Figure 1 and Figure 2 illustrate how varying the missingness threshold ($0.60 \leq \theta \leq 0.95$) affects the number of retained and discarded rows for the 30-minute dataset and 120-minute dataset, respectively. As the threshold becomes more permissive (moving from 0.60 to 0.95):Units

- 30-Minute Data: At the lowest threshold (0.60), only about 1,500 rows are retained and around 5,000 rows are discarded, indicating that many rows exceed 60% missingness. As the threshold increases, stricter row-removal criteria are relaxed, so more rows remain in the dataset. By a threshold of 0.95, nearly 6,500 rows are retained and fewer than 500 are excluded.
- 120-Minute Data: A similar trend appears, although even at $\theta=0.60$ a larger portion of rows (over 4,000) is retained compared to the 30-minute dataset. This suggests that the 120-minute data have fewer high-missingness rows or more complete measurements overall. By $\theta=0.95$, over 6,300 rows remain and only a small fraction are removed.

Overall, these results demonstrate that raising the missingness threshold substantially increases the number of retained samples, especially for the 30-minute data (which generally exhibit higher missing rates). Investigators must balance the desire to keep more data against the risk that excessively incomplete records may introduce noise or reduce imputation quality.

3.2 Best Imputation Method

To identify the optimal approach to handling missing data in both the 30-minute and 120-minute datasets, we systematically varied the row-removal threshold, defined as the maximum allowable proportion of missing features per row, and compared multiple imputation strategies. Performance was assessed using mean squared error

(MSE) and mean absolute error (MAE) relative to a “ground truth” created by artificially masking 10% of the values. For the 120-minute dataset, a threshold of 95% missingness discarded only eight rows while yielding the lowest MSE (968,311) and MAE (91.65) under Iterative Imputer (max_iter=12), demonstrating that even highly incomplete rows could be retained without severely compromising imputation accuracy. By contrast, the 30-minute data benefited from a more stringent threshold of 75%, removing 3,720 rows, to achieve the lowest MSE (962,903) and MAE (91.42) with the same imputation method. Despite differing optimal thresholds, Iterative Imputer consistently outperformed simpler approaches in both time windows, underscoring the utility of modeling inter-variable relationships to achieve robust imputation results.

3.3 Best Machine Learning Model

After applying the various classification algorithms to the 120-minute dataset, Gradient Boosting emerged as the top performer, achieving the highest average F1-score (0.7983) across the cross-validation folds. In terms of overall classification metrics—Accuracy, Precision, Recall, and F1—Gradient Boosting consistently outperformed established ensemble methods such as Random Forest and CatBoost, as well as linear/logistic models (e.g., Logistic Regression) and instance-based methods (KNN). This result indicates that, over a longer 120-minute window of recorded physiological and laboratory parameters, a boosted ensemble approach can successfully capture complex nonlinear patterns that distinguish infected from non-

infected neonates. Figure 3 indicates the top 5 ROC curves in the model selection part for this dataset.

In contrast, the 30-minute data showed CatBoost as the best algorithm by F1-score (0.7634), narrowly surpassing Gradient Boosting (0.7628) and other tree-based methods (e.g., Random Forest, LightGBM) in terms of balanced predictive performance. Although certain algorithms (e.g., Logistic Regression) also demonstrated robust Accuracy, CatBoost's specialized handling of categorical features and iterative boosting led to more balanced gains in Precision and Recall, rendering it particularly effective in this shorter time window, where missing data and limited observation periods can constrain model inputs. Figure 4 shows the top 5 ROC curves in the model selection part for the 30-min dataset.

Notably, the F1-scores for Gradient Boosting (in the 120-minute data) and CatBoost (in the 30-minute data) were sufficiently close that either method could be considered a strong candidate for modeling neonatal infection risk. Consequently, the next stage of experimentation will apply both CatBoost and Gradient Boosting to both time windows under more targeted hyperparameter optimization, ensuring that the final model selection accounts for the subtleties of each approach's performance in different temporal contexts. This two-model strategy aims to solidify which boosting framework ultimately offers the most consistent and clinically relevant predictions.

We conducted an extensive grid search to optimize CatBoost and Gradient Boosting models for both the 30-minute and 120-minute datasets. Each experiment varied key

parameters such as tree depth, number of iterations (or estimators), learning rate, and subsampling rates, with cross-validation (CV) accuracy serving as the primary criterion for ranking configurations. For the 30-minute dataset, CatBoost achieved the highest CV accuracy (approximately 0.7946) under a configuration involving a moderate tree depth (depth=4) and 300 iterations at a learning rate of 0.05. By contrast, the best Gradient Boosting model on the same dataset attained a CV accuracy near 0.7903, reflecting competitive but slightly lower performance relative to CatBoost. Inspection of confusion matrices and classification reports confirmed that these top-ranked parameter sets offered balanced improvements in both Recall and Precision.

In the 120-minute setting, the situation was reversed: Gradient Boosting attained a marginally higher CV accuracy (about 0.8024), surpassing CatBoost's best of approximately 0.8006. Notably, the optimal Gradient Boosting configuration differed from the 30-minute scenario: it favored a shallower tree (max_depth=3) with more estimators (150) and a subsampling rate of 0.8, emphasizing consistent gains in predictive performance across folds. Although both models yielded robust accuracy on the fully trained dataset (ranging between 0.82 and 0.86 when evaluated via confusion matrices), the slight discrepancies between the final accuracy and CV ranking underscore the importance of cross-validation in guiding hyperparameter selection. In particular, the alignment of strong CV scores with high out-of-sample accuracy highlights the capacity of boosted ensemble methods to capture subtle nonlinearities

and interactions within newborn infection data, even under demanding early-time-window conditions.

3.4 Incremental Coverage Analysis

To evaluate how classification performance is influenced by the number of features retained, we performed an incremental coverage analysis where only the top-ranked features (by model-specific importance) were retained in successive subsets. For the 30-minute dataset (using CatBoost), coverage levels ranged from 10% to 80% of the total feature set, with each subset evaluated via 10-fold cross-validation. As shown in the results table, the highest mean accuracy (0.7896) was achieved at 50% coverage, suggesting that roughly half of the most important features were sufficient to obtain robust predictive performance. Notably, while lower coverage (e.g., 10% or 25%) produced a slightly lower accuracy (0.7754 and 0.7815, respectively), including too many features at 75–80% coverage again lowered accuracy, likely due to the introduction of less informative or redundant variables.

A similar procedure was applied to the 120-minute dataset with a Gradient Boosting model, also tested at 10% to 80% coverage. In this case, the peak accuracy (0.8058) occurred at 80% coverage, indicating that a relatively larger proportion of features contributed meaningful signal in the extended 120-minute window. Despite a near-competitive accuracy being observed at 70% coverage (0.8045), the model benefited from retaining a broader set of features, perhaps reflecting the richer data collected over a longer timeframe. Overall, these findings underscore that an optimal balance

exists between too few features (losing important signals) and too many (risking noise or redundancy). By identifying this “sweet spot,” we can streamline model complexity while preserving high predictive performance.

CHAPTER 4: DISCUSSION

In this study, we explored multiple machine learning–based methods for predicting neonatal infection risk and compared their performances with a more conventional reference scoring system. Specifically, we aimed to identify an algorithm that could accurately predict infection-related outcomes when data are aggregated either from the first 30 minutes or the first 120 minutes of intensive care unit (ICU) admission. Our findings revealed notable differences in the models' performances across these two time windows, highlighting the trade-offs of each approach for real-world clinical use.

4.1 Overview of Findings

In the 30-minute dataset, CatBoost emerged as the top-performing model with the highest F1-score (~0.7634), surpassing Gradient Boosting (~0.7628) and other tree-based or linear methods. The marginal gap between CatBoost and Gradient Boosting suggests that, when data are relatively sparse (i.e., at only 30 minutes into an ICU stay), subtle differences in how ensemble algorithms handle missingness, outlier measurements, and imputation steps can substantially shape final performance metrics. CatBoost's specialized handling of categorical data—alongside robust boosting iterations—may have conferred an advantage in the "ultra-early" context, where variable availability is inconsistent, and the physiologic signals can be less stable.

When the time window was extended to 120 minutes, Gradient Boosting models exhibited the highest average F1-score (~0.7983) and displayed excellent calibration.

This improvement likely stems from the greater data richness at 120 minutes; additional laboratory results, updated vital signs, and time for more hemodynamic fluctuations can allow gradient-boosted ensembles to better capture complex, nonlinear relationships. Nonetheless, CatBoost, Random Forest, and certain neural network models also demonstrated strong discriminatory ability (F1-scores typically above 0.76), implying that multiple algorithmic families can achieve clinical usefulness once enough data points become available. The key difference was that Gradient Boosting reached slightly higher and more consistent cross-validation scores across accuracy, precision, and recall metrics.

4.2 Interpretations and Clinical Implications

Our results underscore the importance of time-dependent data availability in infection-risk modeling. While the 30-minute window allows for extremely early intervention, it imposes considerable data limitations, such as fewer blood gas analyses or incomplete laboratory results. As a result, advanced techniques that excel under missing data or can effectively leverage categorical inputs—like CatBoost—may offer a decisive edge in the first 30 minutes.

By contrast, the 120-minute dataset captures a more complete clinical profile; repeated vital-sign measurements, additional laboratory panels, and extended neurological or cardiovascular observations can refine risk estimation. Gradient Boosting's marginally superior performance in the 120-minute window highlights how

ensemble-based approaches can exploit this richer information to model more nuanced interactions among variables.

From a decision-support perspective, these findings suggest a two-stage or adaptive approach: (1) an ultra-early model suitable for the first 30 to 60 minutes of admission, leveraging robust handling of missing data (CatBoost or a similar boosted method), and (2) a refined model (e.g., Gradient Boosting) retrained with updated data around 120 minutes, delivering a more accurate prediction for subsequent critical-care decisions.

4.3 Comparison with Existing Literature

Several prior investigations have compared machine learning models (e.g., Random Forest, XGBoost, and neural networks) against established clinical scoring systems for sepsis or infection-related mortality. Consistent with those reports, we found that ensemble-based algorithms offer better calibration and higher F1-scores than classical logistic regression or basic score systems (e.g., APACHE II or simplified risk scores). Notably, in table2, the incremental gains are especially relevant in the “intermediate” time frame (1–2 hours), when important physiologic trends start to manifest yet remain absent at baseline.

4.4 Limitation

The first one is data quality and heterogeneity. Although we used standardized ICU data from MIMIC-III, variations in measurement frequency, device calibration, and charting practices can introduce biases. The real-world deployment of our model,

especially in other institutions, would require local calibration and external validation. In this study, we explored multiple machine learning–based methods for predicting neonatal infection risk and compared their performances with a more conventional reference scoring system. Specifically, we aimed to identify an algorithm that could accurately predict infection-related outcomes when data are aggregated either from the first 30 minutes or the first 120 minutes of intensive care unit (ICU) admission. Our findings revealed notable differences in the models' performances across these two time windows, highlighting the trade-offs of each approach for real-world clinical use.

We only extracted variables from the first 30 or 120 minutes. Future work could incorporate dynamic, time-series features beyond these discrete windows or investigate sliding-window updates (e.g., every 15 minutes) to capture evolving physiologic patterns.

Although CatBoost and Gradient Boosting can produce feature-importance estimates or SHAP values, the underlying mechanisms remain "black-box" at the bedside. Efforts to embed transparent, clinically interpretable frameworks (e.g., rule-based ensembles) might further facilitate end-user adoption.

APPENDIX

Table 1 Baseline characteristics of the study cohort

	30-minute dataset				120-minute dataset			
Variable	Not infected, mean(std)	Infected, mean(std)	Total, mean(std)	Missing rate (%)	Not infected, mean(std)	Infected, mean(std)	Total, mean(std)	Missing rate (%)
Atypical Lymphocytes	0.80 (1.64)	1.33 (2.66)	1.12 (2.32)	78.84	0.79 (1.59)	1.25 (2.47)	1.03 (2.10)	48.20
BP Cuff [Diastolic]	37.62 (8.16)	33.60 (18.58)	34.77 (16.35)	59.78	37.64 (7.64)	33.41 (14.89)	34.97 (12.86)	25.71
BP Cuff [Mean]	49.59 (8.24)	43.75 (8.83)	45.45 (9.06)	59.78	49.77 (7.61)	43.85 (7.80)	46.04 (8.24)	25.77
BP Cuff [Systolic]	68.97 (9.64)	62.06 (16.46)	64.07 (15.13)	59.78	69.34 (8.99)	61.99 (10.40)	64.71 (10.52)	25.72
Bands	2.54 (3.63)	1.49 (3.00)	1.91 (3.30)	78.83	2.59 (3.79)	1.72 (3.20)	2.15 (3.52)	48.16
Basophils	0.28 (0.60)	0.23 (0.50)	0.25 (0.54)	78.54	0.24 (0.54)	0.23 (0.51)	0.24 (0.52)	47.56
Eosinophils	2.16 (2.25)	2.29 (2.34)	2.24 (2.31)	78.54	1.99 (1.99)	2.24 (2.25)	2.12 (2.13)	47.56
Glucometer	67.60 (20.57)	63.13 (25.64)	64.18 (24.60)	74.77	69.11 (19.12)	69.55 (25.23)	69.43 (23.77)	45.70
Heart Rate	143.93 (16.58)	149.90 (17.15)	148.15 (17.20)	43.71	141.23 (15.17)	146.93 (15.39)	144.83 (15.55)	5.12
Hematocrit	51.67 (5.72)	48.93 (6.65)	50.04 (6.43)	78.35	51.81 (5.71)	49.03 (6.83)	50.38 (6.46)	47.13
Hemoglobin	17.44	16.32	16.78	78.65	17.50	16.38	16.92	47.82

	(1.93)	(2.16)	(2.14)		(1.92)	(2.25)	(2.17)	
Lymphocytes	30.40 (13.73)	49.31 (19.32)	41.69 (19.62)	78.54	28.22 (12.17)	46.79 (19.92)	37.79 (19.03)	47.56
MCH	35.56 (1.94)	36.59 (2.61)	36.17 (2.41)	78.65	35.56 (1.89)	36.74 (2.46)	36.17 (2.28)	47.82
MCHC	33.76 (0.94)	33.36 (1.09)	33.52 (1.05)	78.67	33.78 (0.91)	33.45 (1.06)	33.61 (1.00)	47.86
MCV	105.39 (5.27)	109.80 (8.17)	108.01 (7.45)	78.67	105.36 (5.31)	109.96 (7.71)	107.72 (7.04)	47.86
Metamyelocytes	0.25 (0.68)	0.22 (0.78)	0.23 (0.74)	78.90	0.27 (0.70)	0.21 (0.72)	0.24 (0.71)	48.47
Monocytes	6.87 (3.73)	6.78 (3.65)	6.82 (3.68)	78.54	7.11 (3.63)	6.91 (3.82)	7.01 (3.73)	47.56
Myelocytes	0.15 (0.68)	0.12 (0.55)	0.13 (0.60)	78.96	0.14 (0.56)	0.12 (0.52)	0.13 (0.54)	48.48
Neutrophils	56.59 (14.78)	38.07 (18.80)	45.53 (19.53)	78.54	58.66 (13.49)	40.41 (19.34)	49.26 (19.08)	47.56
Platelet Count	307.43 (73.73)	275.28 (80.91)	288.21 (79.65)	78.64	301.87 (78.21)	265.36 (81.77)	283.00 (82.11)	47.68
RDW	16.62 (0.97)	17.17 (1.41)	16.95 (1.28)	78.68	16.65 (1.01)	17.13 (1.37)	16.90 (1.23)	47.94
Red Blood Cells	4.92 (0.59)	4.48 (0.71)	4.65 (0.70)	78.67	4.93 (0.57)	4.47 (0.70)	4.70 (0.68)	47.86
Resp Rate	49.70 (13.33)	51.90 (14.84)	51.25 (14.44)	44.55	48.11 (11.37)	52.17 (12.84)	50.66 (12.47)	6.29
SaO2	97.48 (5.04)	96.00 (4.49)	96.35 (4.67)	49.92	97.85 (3.92)	96.33 (3.16)	96.77 (3.47)	19.10
Temp Axilary [F]	98.41 (0.70)	98.51 (0.83)	98.49 (0.80)	83.25	98.52 (0.61)	98.71 (0.71)	98.66 (0.69)	47.51

Temp Skin [C]	36.60 (4.77)	36.57 (3.94)	36.57 (4.10)	71.92	36.53 (3.94)	36.58 (3.72)	36.57 (3.77)	41.33
Temp/Iso/Warmer	35.95 (3.02)	36.27 (1.84)	36.20 (2.15)	66.63	35.87 (3.05)	36.13 (2.36)	36.06 (2.55)	33.15
White Blood Cells	17.06 (5.57)	13.00 (6.01)	14.64 (6.16)	78.42	17.27 (5.47)	13.02 (6.54)	15.08 (6.41)	47.32

Table 2: Comparison between Different Infection Detection Methods

Detection Method	Time Window	Accuracy(%)
SNAP-II Score [37]	12-24 hours	~ 75
CRIB Score [38]	12-24 hours	~ 73
Naïve Bayes [39]	12 hour	~ 78
New model	30 – 120 min	~ 80

Figure 1: Research Pipeline

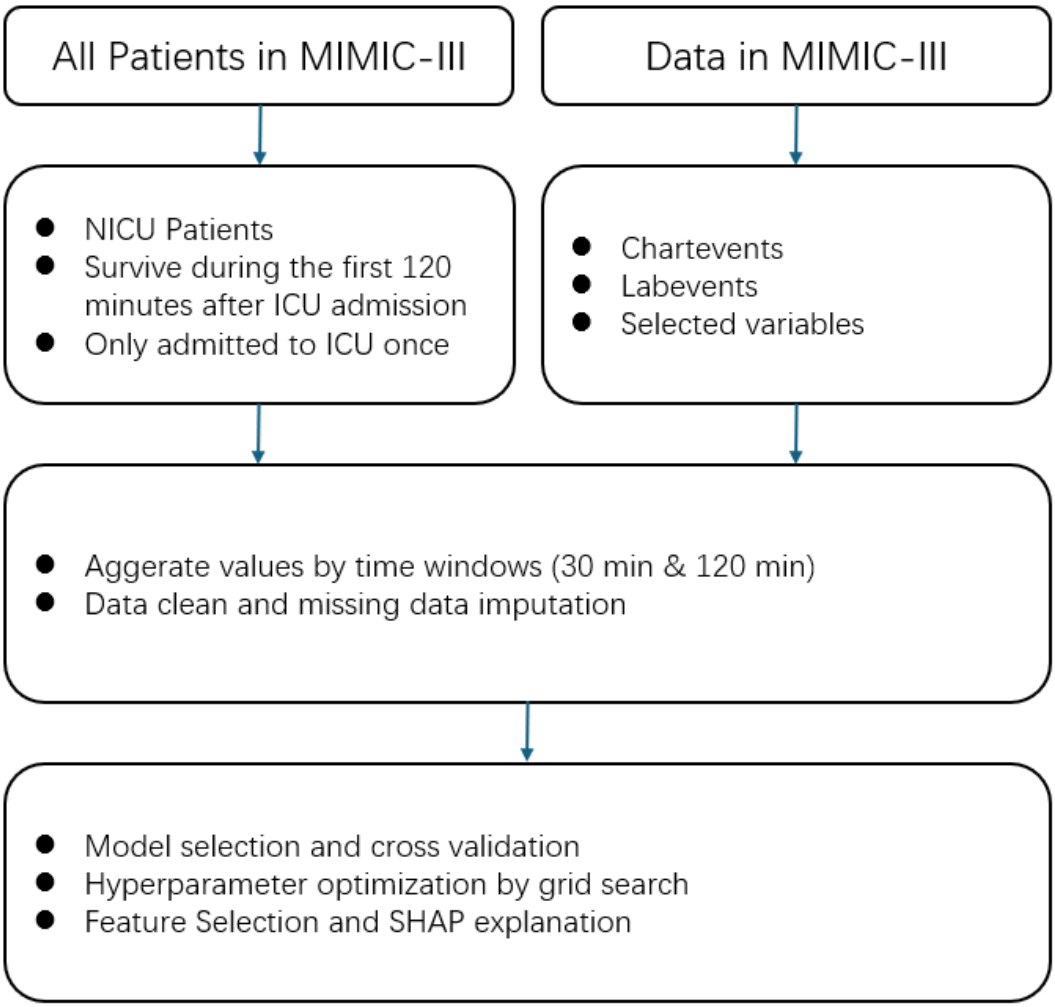


Figure 2: Missing Value Distribution

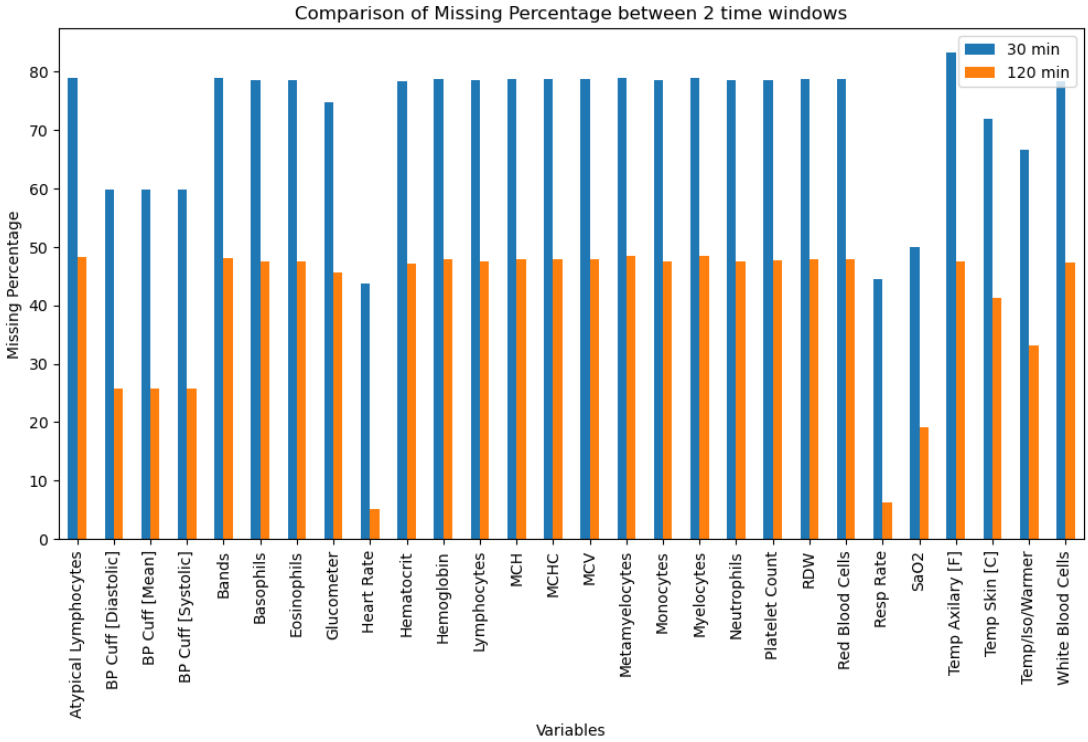


Figure 3: Top Five ROC Curve in Model Selection of 120 Minutes Dataset

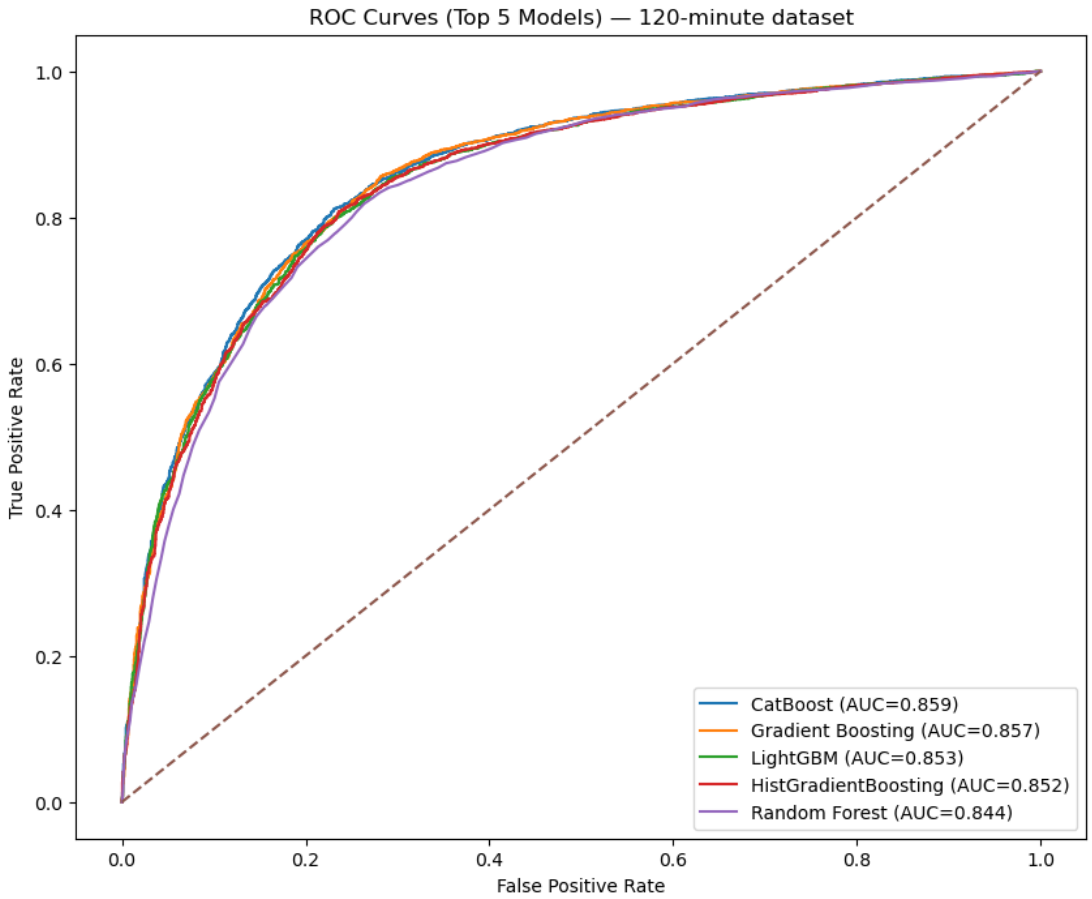
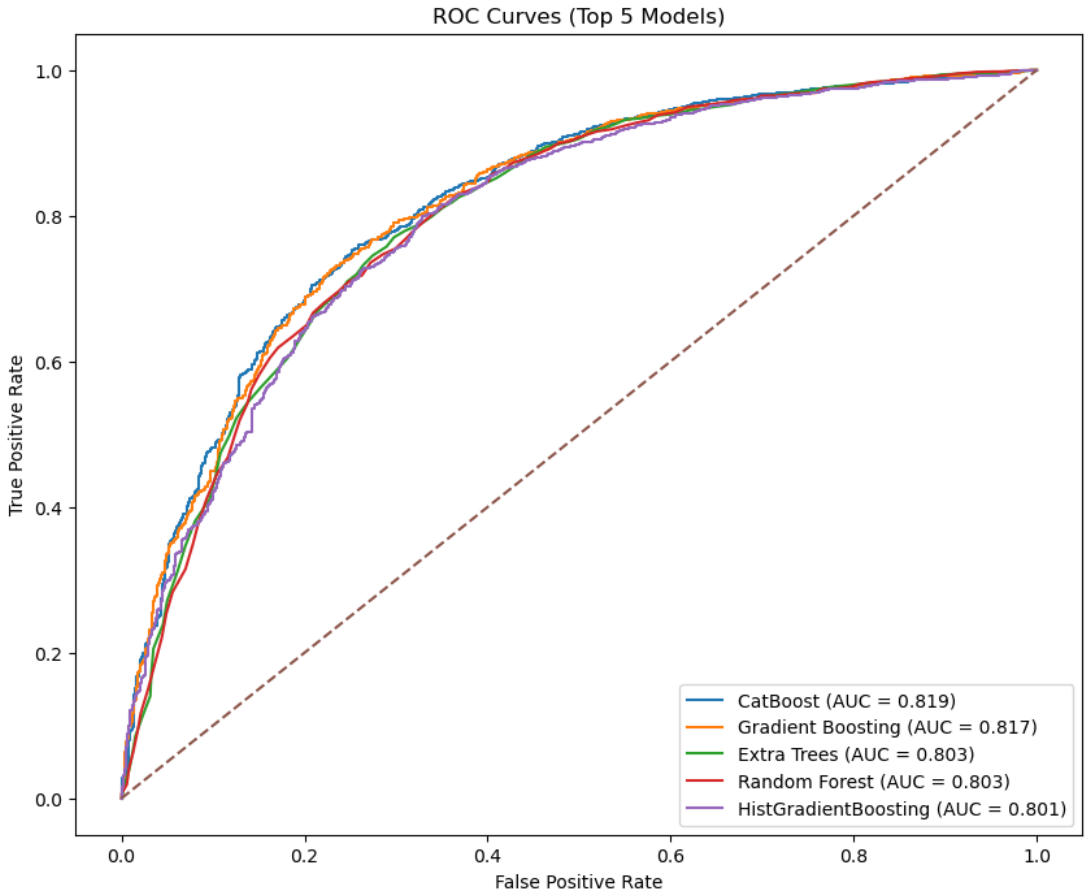


Figure 4: Top Five ROC Curve in Model Selection of 120 Minutes Dataset



REFERENCE

1. Seale AC, Blencowe H, Manu AA, Nair H, Bahl R, Qazi SA, Zaidi AK, Berkley JA, Cousens SN, Lawn JE., pSBI Investigator Group. Estimates of possible severe bacterial infection in neonates in sub-Saharan Africa, south Asia, and Latin America for 2012: a systematic review and meta-analysis. Lancet Infect Dis. 2014 Aug;14(8):731-741.
2. Fleischmann-Struzek C, Goldfarb DM, Schlattmann P, Schlapbach LJ, Reinhart K, Kissoon N. The global burden of paediatric and neonatal sepsis: a systematic review. Lancet Respir Med. 2018 Mar;6(3):223-230. doi: 10.1016/S2213-2600(18)30063-8. PMID: 29508706.
3. Stoll, B. J., Hansen, N. I., Bell, E. F., Walsh, M. C., Carlo, W. A., Shankaran, S., ... & Eunice Kennedy Shriver National Institute of Child Health and Human Development Neonatal Research Network.
4. Cantey JB, Patel SJ. Antimicrobial stewardship in the NICU. Infect Dis Clin North Am. 2014 Jun;28(2):247-61. doi: 10.1016/j.idc.2014.01.005. PMID: 24857391.
5. Richardson DK, Corcoran JD, Escobar GJ, Lee SK. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. J Pediatr. 2001 Jan;138(1):92-100. doi: 10.1067/mpd.2001.109608. PMID: 11148519.
6. Parry G, Tucker J, Tarnow-Mordi W; UK Neonatal Staffing Study Collaborative Group. CRIB II: an update of the clinical risk index for babies score. Lancet. 2003 May 24;361(9371):1789-91. doi: 10.1016/S0140-6736(03)13397-1. PMID: 12781540.

7. Parry G, Tucker J, Tarnow-Mordi W; UK Neonatal Staffing Study Collaborative Group. CRIB II: an update of the clinical risk index for babies score. Lancet. 2003 May 24;361(9371):1789-91. doi: 10.1016/S0140-6736(03)13397-1. PMID: 12781540.

8. Fairchild KD, O'Shea TM. Heart rate characteristics: physiomarkers for detection of late-onset neonatal sepsis. Clin Perinatol. 2010 Sep;37(3):581-98. doi: 10.1016/j.clp.2010.06.002. PMID: 20813272; PMCID: PMC2933427.

9. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019 Apr 4;380(14):1347-1358. doi: 10.1056/NEJMra1814259. PMID: 30943338.

10. Mani S, Ozdas A, Aliferis C, Varol HA, Chen Q, Carnevale R, Chen Y, Romano-Keeler J, Nian H, Weitkamp JH. Medical decision support using machine learning for early detection of late-onset neonatal sepsis. J Am Med Inform Assoc. 2014 Mar-Apr;21(2):326-36. doi: 10.1136/amiajnl-2013-001854. Epub 2013 Sep 16. PMID: 24043317; PMCID: PMC3932458.

11. Baker S, Kandasamy Y. Machine learning for understanding and predicting neurodevelopmental outcomes in premature infants: a systematic review. Pediatr Res. 2023 Jan;93(2):293-299. doi: 10.1038/s41390-022-02120-w. Epub 2022 May 31. PMID: 35641551; PMCID: PMC9153218.

12. Johnson, A., Pollard, T., Shen, L. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016). https://doi.org/10.1038/sdata.2016.35

13. Harutyunyan, H., Khachatrian, H., Kale, D.C. *et al.* Multitask learning and benchmarking with clinical time series data. *Sci Data* **6**, 96 (2019). https://doi.org/10.1038/s41597-019-0103-9

14. Ahmad, M. A., Teredesai, A., & Eckert, C. (2018). Interpretable Machine Learning in Healthcare. Proceedings of the IEEE, 106(11), 2022-2038.

15. Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised learning. International Journal of Computer Science, 1(2), 111-117.

16. García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge-Based Systems, 98, 1-29.

17. Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. Scientific Data, 6, 96.

18. Polin RA; Committee on Fetus and Newborn. Management of neonates with suspected or proven early-onset bacterial sepsis. Pediatrics. 2012 May;129(5):1006-15. doi: 10.1542/peds.2012-0541. Epub 2012 Apr 30. PMID: 22547779.

19. Polin RA, Carlo WA; Committee on Fetus and Newborn; American Academy of Pediatrics. Surfactant replacement therapy for preterm and term neonates with respiratory distress. Pediatrics. 2014 Jan;133(1):156-63. doi: 10.1542/peds.2013-3443. Epub 2013 Dec 30. PMID: 24379227.

20. Shane AL, Sánchez PJ, Stoll BJ. Neonatal sepsis. Lancet. 2017 Oct 14;390(10104):1770-1780. doi: 10.1016/S0140-6736(17)31002-4. Epub 2017 Apr 20. PMID: 28434651.

21. Goldstein B, Giroir B, Randolph A; International Consensus Conference on Pediatric Sepsis. International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics. Pediatr Crit Care Med. 2005 Jan;6(1):2-8. doi: 10.1097/01.PCC.0000149131.72248.E6. PMID: 15636651.

22. Hornik CP, Benjamin DK, Becker KC, Benjamin DK Jr, Li J, Clark RH, Cohen-Wolkowiez M, Smith PB. Use of the complete blood cell count in early-onset neonatal sepsis. Pediatr Infect Dis J. 2012 Aug;31(8):799-802. doi: 10.1097/INF.0b013e318256905c. PMID: 22531231; PMCID: PMC3399972.

23. Polin RA; Committee on Fetus and Newborn. Management of neonates with suspected or proven early-onset bacterial sepsis. Pediatrics. 2012 May;129(5):1006-15. doi: 10.1542/peds.2012-0541. Epub 2012 Apr 30. PMID: 22547779.

24. Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

25. Sterne, J. A., et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ, 338, b2393.

26. Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data. John Wiley & Sons.

27. Bilogur, A. (2018). Missingno: a missing data visualization suite. Journal of Open Source Software, 3(22), 547.

28. Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404), 1198-1202.

29. Enders, C. K. (2010). Applied missing data analysis. Guilford Press.

30. Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5-6), 519-533.

31. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45, 1-67.

32. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

33. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI (Vol. 14, No. 2, pp. 1137-1145).

34. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.

35. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(10), 281-305.

36. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

37. Richardson DK, Corcoran JD, Escobar GJ, Lee SK. SNAP-II and SNAPPE-II: Simplified newborn illness severity and mortality risk scores. J Pediatr. 2001 Jan;138(1):92-100. doi: 10.1067/mpd.2001.109608. PMID: 11148519.

38. Parry, G., Tucker, J., & Tarnow-Mordi, W. (2003). CRIB II: an update of the clinical risk index for babies score. The Lancet, 361(9371), 1789-1791.

39. Mani, S., Ozdas, A., Aliferis, C., Varol, H. A., Chen, Q., Carnevale, R., ... & Weitkamp, J. H. (2014). Medical decision support using machine learning for early detection of late-

onset neonatal sepsis. Journal of the American Medical Informatics Association, 21(2), 326-336.