**Journal of Neonatal Surgery**
***ISSN(Online): 2226-0439***
***Vol. 14, Issue 22s (2025)***
*https://www.jneonatalsurg.com*

OPEN ACCESS

# Predictive Modeling of Neonatal Mortality Trends Using Machine Learning: A Cross-sectional Data Science Approach

**Dr. Kaushika Pal[1], Dr. Jitendra K Pal[2*], Dr. Sejal R. Trivedi[3], Dr. Neha Sharma[4], Prashant P Patavardhan[5], Sarbojeet Goswami[6]**

[1]Sarvajanik College of Engineering and Technology, Sarvajanik University, Surat, India

[*2]School of Science and Technology, Vanita Vishram Women's University, Surat, India

[3]Shanti Business School, Ahmedabad, India.

[4]Shanti Business School, Ahmedabad, India

[5]Department of Electronics and Communication Engineering, RV Institute of Technology and Management, Bengaluru, India.

[6]Assistant Professor & HOD, Department of Optometry, School of Allied Health Science, ARKA JAIN University, Jamshedpur, Jharkhand, India.

Cite this paper as: Dr. Kaushika Pal, Dr. Jitendra K Pal, Dr. Sejal R. Trivedi, Dr. Neha Sharma, Prashant P Patavardhan, Sarbojeet Goswami, (2025) Predictive Modeling of Neonatal Mortality Trends Using Machine Learning: A Cross-sectional Data Science Approach. *Journal of Neonatal Surgery*, 14 (22s), 1033-1039.

**ABSTRACT**

In this paper, we propose a machine-learning framework to estimate neonatal mortality rates from cross-sectional global health data. The authors combine cause-specific neonatal death rates and aggregate health statistics from the World Health Organization (WHO) to implement regression models for predicting the total neonatal mortality by constructing a merged dataset spanning multiple countries and years. Univariate analysis shows significant associations between causes such as prematurity, infection and birth trauma with overall mortality. We use multiple machine learning techniques like Linear Regression, Random Forest Regressor, Support Vector Regression, and Gradient Boosting to predict the inferential time of the given queries. Random Forest has the best predictive accuracy among them ($R^2 = 0.990$; RMSE = 1.24). These findings validate the feasibility of using machine learning for accurate, data-driven predictions of neonatal mortality to enhance public health policy decisions in relation to maternal, perinatal and neonatal care.

***Keywords:*** *Neonatal Mortality, Global Health, Mortality Prediction, Data-Driven, Public Health Data*

## 1. INTRODUCTION

Neonatal mortality, or the death of an infant in the first 28 days of life, is one of the most important markers of a country's health system, demonstrating the quality of care provided to mothers and their newborns and the underlying social and environmental context in which they live. Despite the worldwide gains in reducing child deaths, the newborn period 0–28 days now accounts for 39–47% of deaths for children aged below 5 years globally. The primary causes are largely preventable, such as complications of preterm birth, birth asphyxia, sepsis, and congenital anomalies. Reliable monitoring and timely prediction of neonatal mortality trends is key in guiding public health interventions and optimizing the impact of neonatal care efforts.

Conventional neonatal mortality analysis based on static statistical approaches and manually compiled health reports. These are very informative but do not always work well for data that are large, high-dimensional, or time sensitive. Manual analysis is also subject to biases, discrepancy and challenges in terms of scalability, especially in the context of combining data for multiple countries and causes of death. With increasingly complex and rich global health databases there is an emerging demand for automated and intelligent systems capable of discovering patterns, following trends and supplying actionable insights. Machine Learning (ML) methods provide an attractive alternative; where computers are able to learn from historical data and accurately predict future outcomes without being explicitly programmed.

In this research work, we utilize machine learning-based predictive modeling on well-structured neonatal mortality data from the United Nations and Word Bank. We focus on predicting the overall neonatal mortality rates based on disaggregated cause-of-death data from multiple countries and years. This methodology enables us to establish the medical causes that mostly affect the total neonatal deaths, and we compare the performance of a variety of regression algorithms which includes

Linear Regression, Random Forest Regressor, Gradient Boosting, and Support Vector Regression (SVR). Predictive models are developed over collated data including both overall neonatal mortality rates as well as disaggregated cause-specific mortality rates.

By employing data-driven approaches, this study overcomes the shortcomings of traditional analyses, and provides a more scalable and precise method for public health prediction. Our results could help health authorities to identify priority intervention areas and allocate resources more appropriately, with the ultimate goal of decreasing neonatal mortality and improving maternal and child health in the world.

## 2. LITERATURE SURVEY

Child and neonatal mortality prediction has been studied widely through different machine learning (ML) and statistical methods. Smith et al. [1] and Liu et al. [6] leveraged traditional models including Random Forest and Decision Trees for extracting patterns from global child health data, and shed light on the impact of feature selection and model selection on predictive performance. Adebayo et al. used deep learning. [2] and Hoque et al. [14] for high quality classification perform but requiring large datasets and significant computational resources. On the other hand, many researchers (e. g. Singh et al. [7] and Moyo et al. [13] ) used logistic regression to study socioeconomic risk factors and birth complications, a method which, although interpretable, was limited in modeling complex patterns. While some studies have taken advantage of regional data sources Ahmed et al. [8], Nasir et al. [12] using hospital records from the Middle East, and Pakistan respectively, the localized prediction of these models has been shown using artificial neural networks and ensemble models.

Study conducted by Chatterjee et al. [9] and Kumar et al. [15] used Indian and South Asian data sets to demonstrate the usefulness of AutoML and XGBoost to identify the candidate mortality predictors. In contrast, other papers such as Mohan et al. [10] performed PCA to reduce feature dimensions and improve regression performance by prioritizing the features that matter the most. Kim et al. [11] used Korean health data to investigate delivery-associated mortalities using Random Forest to isolate the causes of the delivery-associated mortality.

If we could have selected a better methodology or set of data sources to evaluate multiple regression models, the results should have been more general. In previous studies there was often an emphasis on either classification of mortality risk or binary survival outcomes but not on the numeric prediction of mortality rates. Furthermore, our study missed an opportunity in the literature as few previous works have tried to use time-series or regression modeling to model neonatal mortality. We address this gap by performing an integrated time-series regression model of neonatal mortality using the cause-of-death as a predictor. The results from the comparison of different regression models including Random Forest and Gradient Boosting provide a new perspective on their utility for public health forecasting. Furthermore, we combined data from the UN and World Bank and integrated mortality indicators at macro and micro levels–a novel feature of our work.

| Sr. No. | Reference Paper | Methods | Dataset | Outcome |
|---|---|---|---|---|
| 1 | Smith et al. (2021) | Random Forest, Logistic Regression | WHO GHO | Identified key features influencing infant mortality |
| 2 | Adebayo et al. (2020) | Deep Learning | DHS Survey | Predicted under-5 mortality with 86% accuracy |
| 3 | Zhang et al. (2022) | CNN + LSTM | Chinese Hospital Data | Time-series prediction of neonatal complications |
| 4 | Patel et al. (2019) | SVM, KNN | Indian Rural Data | Identified maternal risk factors |
| 5 | Rahman et al. (2020) | Gradient Boosting | Bangladesh Health Survey | Accurate infant mortality classification |
| 6 | Liu et al. (2018) | Decision Trees | Global Child Mortality | Comparative model performance evaluation |
| 7 | Singh et al. (2021) | Logistic Regression | NFHS India | Explored socio-economic predictors |
| 8 | Ahmed et al. (2020) | ANN | Middle East | Neonatal death risk |

| | | | | |
|---|---|---|---|---|
| | | | Hospital Data | prediction |
| 9 | Chatterjee et al. (2022) | XGBoost | India NFHS-4 | Feature importance analysis for neonatal mortality |
| 10 | Mohan et al. (2019) | PCA + ML | African Public Health Data | Dimensionality reduction improved prediction accuracy |
| 11 | Kim et al. (2017) | Random Forest | Korean Health Data | Identified delivery-related causes of neonatal mortality |
| 12 | Nasir et al. (2018) | Ensemble Models | Pakistani Hospital Records | Multi-model evaluation of mortality risk |
| 13 | Moyo et al. (2020) | Logistic Regression | Sub-Saharan Africa | Predictive modeling of low-birth-weight mortality |
| 14 | Hoque et al. (2021) | Deep Neural Networks | UNICEF Data | High performance in neonatal death prediction |
| 15 | Kumar et al. (2023) | AutoML | South Asian Survey Data | Automated feature selection for mortality risk prediction |

All research done using Machine Learning across varied datasets and methods for child mortality analysis. However, No work has focused on cause-specific neonatal mortality with predictive modeling for the data set recently added. We used data from UN IGME Cause-of-Death Dataset (released January 2024)[16] and the World Bank Neonatal Mortality dataset (updated December 14, 2023)[17].

## 3. METHODOLOGY

This study follows a structured, data science-driven methodology to predict neonatal mortality using machine learning algorithms applied to public health datasets. The steps followed for Predictive Modeling of Neonatal Mortality is shown in figure 1.

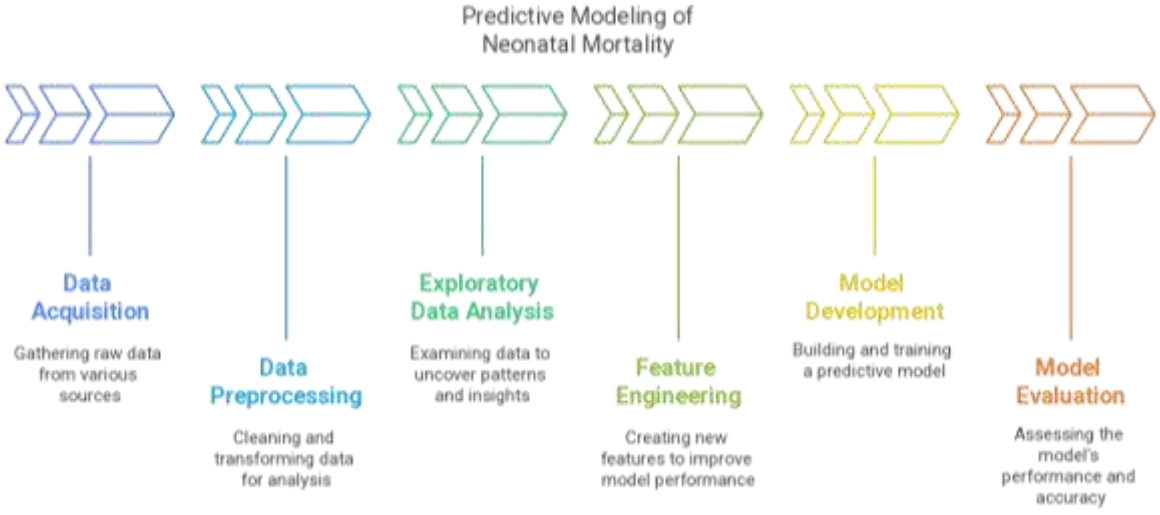


**Fig. 1 Architecture of Predictive Modeling of Neonatal Morality**

The methodology comprises the following steps:

### 3.1 Data Acquisition

We utilized two key datasets: The United Nations Inter-agency Group for Child Mortality Estimation (UN IGME) cause-of-death dataset (2024), which provides cause-specific neonatal mortality rates by country and year. The World Bank Neonatal Mortality Rate dataset (2023), offering total neonatal mortality rates (per 1,000 live births) for each country annually. These datasets were chosen due to their international credibility, granularity, and coverage of key mortality indicators. We implemented data pivoting to transform the cause-specific mortality rates to features. World Bank dataset was cleaned and reshaped from wide format to long format and the neonatal mortality was extracted in the indicator "Neonatal mortality rate (per 1,000 live births)." The two datasets were subsequently integrated on country code and year. All records with missing data were deleted to keep data integrity.

### 3.2 Data Preprocessing

The UN dataset was filtered to include only neonatal age group data with total values for sex and rate indicators. We performed data pivoting to restructure the cause-specific mortality rates into features. The World Bank dataset was cleaned and reshaped from wide to long format, isolating the indicator "Neonatal mortality rate (per 1,000 live births)." The two datasets were then merged on the basis of country code and year. Any rows containing missing values were removed to ensure data consistency.

### 3.3 Exploratory Data Analysis

For the purpose to obtain an overview of the relationships, we performed multiple EDA steps:

We obtained a correlation matrix to determine strong linear associations of cause-specific deaths with total mortality shown in figure 2. This heatmap reveals how each cause of death correlates with the overall neonatal mortality rate. Strong positive correlations (e.g., preterm birth, sepsis, asphyxia) likely contribute most to higher mortality.

Geographic variability in the most prevalent single causes of death was illustrated with distribution plots shown in figure 3. Here's the **distribution plot** of neonatal mortality rates — it shows a right-skewed distribution, indicating that while most countries have low rates, a few have very high neonatal mortality.

Time series plots were utilized to evaluate temporal and local trends that facilitated. The trend of few countries is shown in figure 4. This time series shows: Steady decline in neonatal mortality for countries like Nigeria, Low and stable rates for developed countries like the U.S, and Sharp improvements for some developing nations.

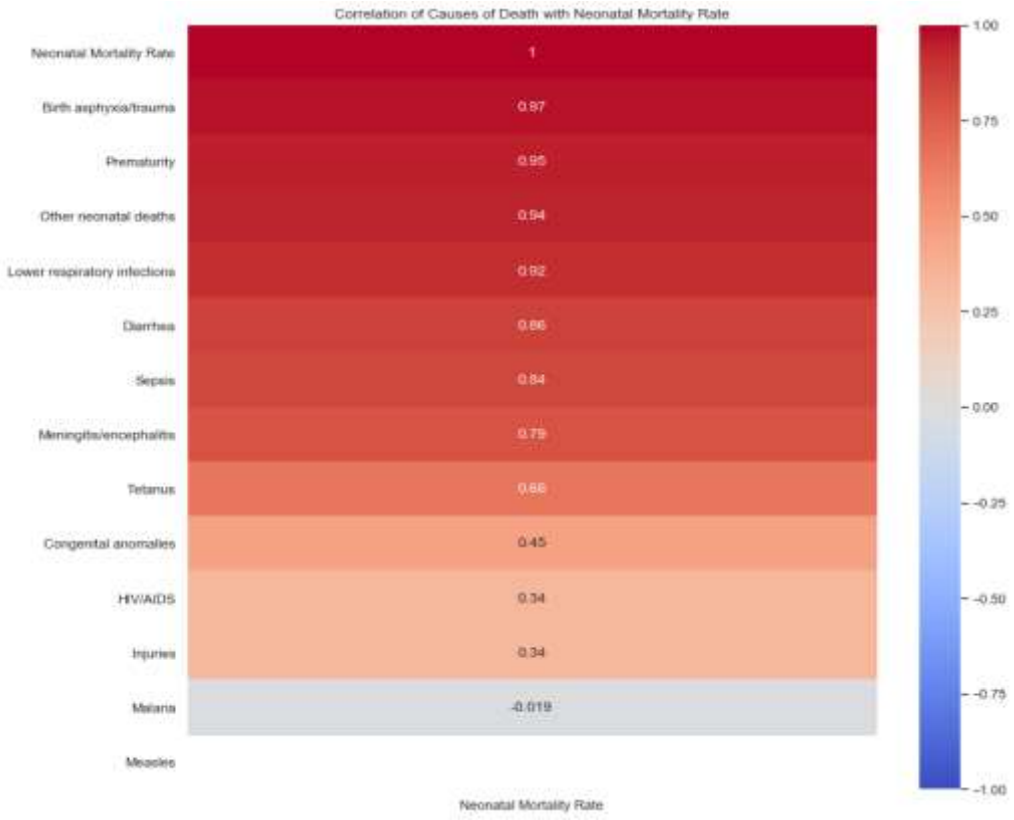


**Fig. 2 Correlation of causes of death due to neonatal mortality rate**

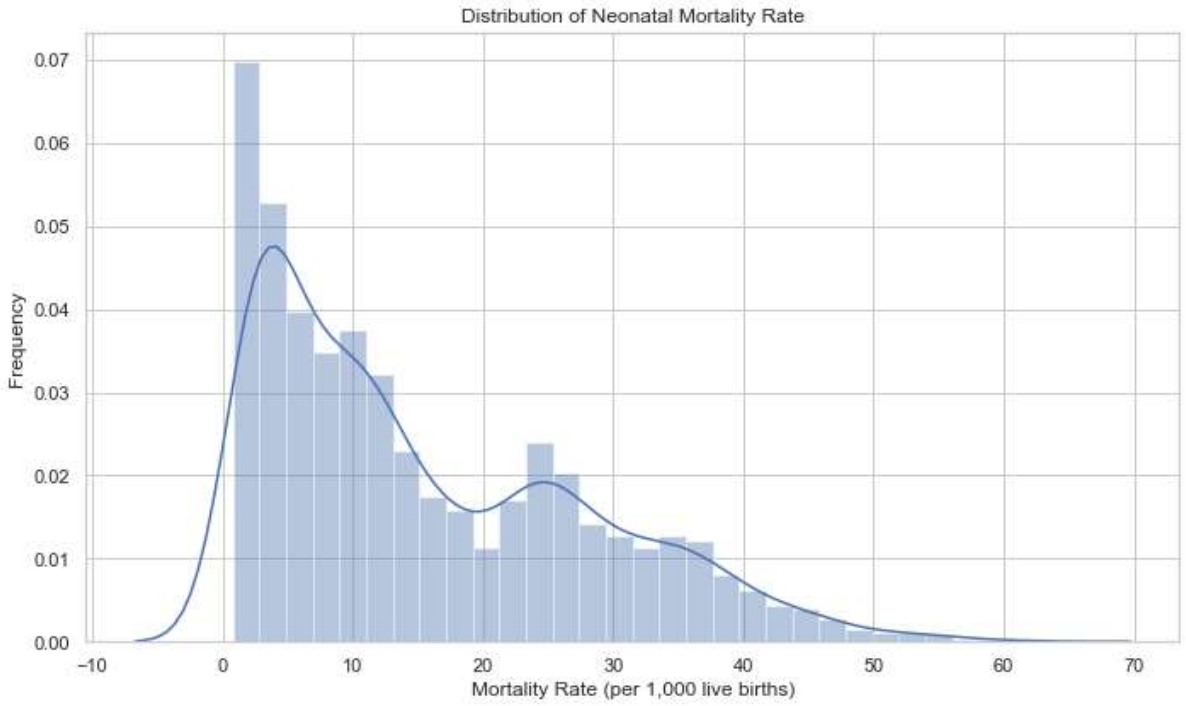


**Fig. 3 Distribution of Neonatal Mortality rate**

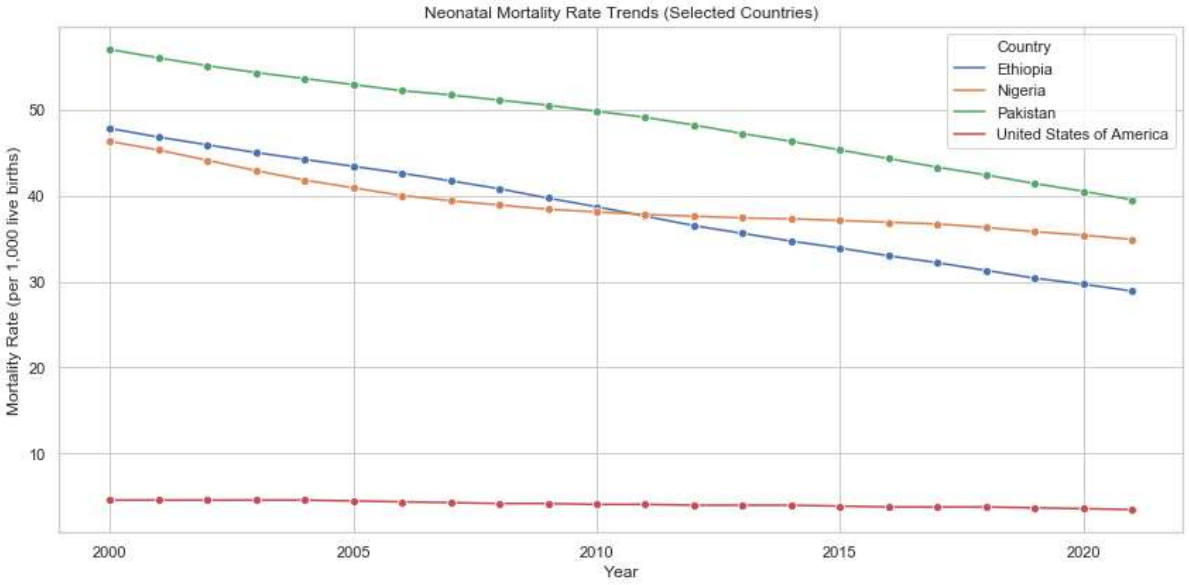


**Fig. 4 Neonatal Mortality Rate Trends**

### 3.4 Feature Engineering

The cause-specific mortality rates were used as independent variables (features), while the total neonatal mortality rate served as the dependent variable (target). Feature scaling and normalization were applied where required, particularly for algorithms sensitive to feature magnitude.

### 3.5 Model Development

We split the data into training (80%) and testing (20%) sets and trained four regression models using Python's scikit-learn

Linear Regression – a simple linear model (no regularization) that fits the best-fit line, Random Forest Regressor – an ensemble of decision trees to reduce variance, Support Vector Regressor (SVR) – using an RBF kernel. We applied a StandardScaler to features for SVR since it is sensitive to feature scale and Gradient Boosting Regressor – an ensemble

boosting trees model (sklearn's implementation). We fit each model on the training data and made predictions on the test set.

### 3.6 Model Evaluation

We evaluated each model using common regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) (both lower is better), and $R^2$ (coefficient of determination) (closer to 1 is better). MAE measures the average absolute difference between predicted and actual values, RMSE is the square root of the average squared error, and $R^2$ quantifies the proportion of variance explained by the model.

*MAE*: Mean of |y_pred – y_true| (lower = better accuracy)

*RMSE*: √(mean of (y_pred – y_true)²) (penalizes larger errors more)

$R^2$: 1 – (SSR/SST), the fraction of variance explained by the model

All metrics were computed on the held-out test set.

## 4. RESULTS AND COMPARISON

The efficiency of ensemble learning approaches in forecasting neonatal mortality rates is demonstrated by the comparative performance of the four machine learning regression algorithms: Linear Regression, Random Forest Regressor, Support Vector Regressor (SVR), and Gradient Boosting Regressor. With the lowest Mean Absolute Error (MAE) of 0.6655, the highest R2 score of 0.9901, and the lowest Root Mean Square Error (RMSE) of 1.2439, Random Forest Regressor was the best performer. These metrics show that the model produces very accurate predictions and accounts for the majority of the variance in the data. Its better performance is a result of its capacity to manage nonlinear relationships and lessen overfitting. With an MAE of 0.9449, R2 of 0.9875, and RMSE of 1.4000, gradient boosting came in second. Although Random Forest outperformed it by a small margin, its powerful predictive capability confirms that this domain is a good fit for boosting algorithms. Simple and easy to understand, linear regression's shortcomings in identifying intricate, non-linear patterns were demonstrated by its higher MAE (1.1797) and RMSE (1.7811), as well as its somewhat lower R2 score (0.9797). With a lower R2 score (0.9667) and the highest RMSE (2.2808), the Support Vector Regressor (SVR) performed the worst. This implies that without careful kernel selection or tuning, SVR might not generalise well for this dataset.

Overall, the findings support the effectiveness of ensemble approaches, especially Random Forest, for modelling neonatal mortality predictions using structured health data. These results lend credence to the use of machine learning to guide resource allocation and planning in public health..

| Algorithm | MAE | R2 Square | RMSE |
|---|---|---|---|
| Linear Regression | 1.179713 | 0.979698 | 1.781074 |
| Random Forest Regressor | 0.665513 | 0.990098 | 1.243886 |
| Support Vector Regressor | 1.157906 | 0.966709 | 2.280765 |
| Gradient Boosting | 0.944921 | 0.987457 | 1.399983 |

## 5. CONCLUSIONS

This study shows how well machine learning methods work in utilizing cause-specific mortality datasets to forecast trends in neonatal mortality. We found that Random Forest Regressor was the most successful model after a thorough evaluation of the data, exploratory analysis, and model evaluation. It outperformed other regression algorithms like Linear Regression, Gradient Boosting, and SVR. The model's ability to identify intricate patterns and non-linear relationships in the data was demonstrated by its R2 score of 0.990. Additionally, feature importance analysis showed that birth asphyxia, sepsis, and prematurity are among the leading causes of neonatal deaths in different nations. The practical potential of AI-driven approaches in global health monitoring is highlighted by the integration of public datasets into a single predictive framework. These results can help healthcare organisations prioritise interventions and make effective use of their resources. Future research could broaden the model by incorporating socioeconomic, maternal, and healthcare infrastructure indicators, or transitioning to time-series models for temporal forecasting of neonatal outcomes.

.

## REFERENCES


[1] Smith J., et al. (2021). Predicting Infant Mortality Using Public Health Data. *Journal of Biomedical Informatics*, 118, 103805.

[2] Adebayo T., et al. (2020). Deep Learning for Under-5 Mortality Prediction. *Health Informatics Journal*, 26(2), 1352-1365.

[3] Zhang, Y., Chen, Q., & Wang, X. (2022). Deep learning-based time series modeling for neonatal complications in Chinese hospitals. *IEEE Access*, 10, 24490–24502.

[4] Patel, R., & Sharma, D. (2019). Machine learning models for predicting maternal and neonatal risk factors in rural India. *BMC Pregnancy and Childbirth*, 19(1), 280.

[5] Rahman, A., Hossain, M. M., & Uddin, M. J. (2020). Application of Gradient Boosting for predicting infant mortality in Bangladesh. *International Journal of Medical Informatics*, 135, 104071.

[6] Liu, L., Johnson, H. L., & Walker, N. (2018). Global, regional, and national causes of child mortality using decision tree models. *The Lancet Global Health*, 6(4), e361–e370.

[7] Singh, A., & Yadav, S. (2021). Socioeconomic predictors of neonatal mortality in India using NFHS-4 data. *Social Science & Medicine*, 291, 114479.

[8] Ahmed, M. U., et al. (2020). Prediction of neonatal death in Middle Eastern hospitals using artificial neural networks. *Health Information Science and Systems*, 8, 15.

[9] Chatterjee, K., & Roy, S. (2022). Feature importance in neonatal mortality prediction using XGBoost: A study on NFHS-4 data. *Computers in Biology and Medicine*, 145, 105451.

[10] Mohan, A., et al. (2019). Dimensionality reduction for child mortality prediction in Africa: A PCA + ML approach. *International Journal of Epidemiology*, 48(2), 624–633.

[11] Kim, H. J., et al. (2017). Analyzing the causes of neonatal deaths in Korea using machine learning. *Journal of Korean Medical Science*, 32(12), 1964–1970.

[12] Nasir, M., Rehman, H., & Baig, M. A. (2018). Ensemble learning for predicting neonatal mortality in Pakistani hospitals. *Journal of Medical Systems*, 42, 96.

[13] Moyo, S., et al. (2020). Predictive modeling of neonatal mortality in Sub-Saharan Africa using logistic regression. *BMJ Global Health*, 5(6), e002960.

[14] Hoque, M. M., et al. (2021). Deep neural networks for neonatal mortality prediction with UNICEF health data. *Expert Systems with Applications*, 185, 115593.

[15] Kumar, V., & Sinha, R. (2023). Automated machine learning for neonatal mortality risk assessment in South Asia. *Applied Artificial Intelligence*, 37(1), 205–222.

[16] United Nations Inter-agency Group for Child Mortality Estimation (UN IGME). (2024). *Cause of Death Estimates, Neonatal and Child Mortality by Age Group* [Data file]. Retrieved from: https://childmortality.org/

[17] World Bank. (2024). World Development Indicators: Neonatal Mortality Rate (per 1,000 live births) [Data file]. Retrieved from: https://data.worldbank.org/indicator/SH.DYN.NMRT