scientific reports

Check for updates

OPEN

# Development and evaluation of machine learning training strategies for neonatal mortality prediction using multicountry data

Gabriel Ferreira dos Santos Silva[1]✉, Roberta Moreira Wichmann[2,3], Francisco Costa da Silva Junior[3] & Alexandre Dias Porto Chiavegatto Filho[1]

**Neonatal mortality poses a critical challenge in global health, particularly in low- and middle-income countries. Leveraging advancements in technology, such as machine learning (ML) algorithms, offers the potential to improve neonatal care by enabling precise prediction and prevention of mortality risks. This study utilized the Maternal and Neonatal Health Registry (MNHR) dataset from the National Institutes of Health (NIH), encompassing multicentric neonatal data across various countries, to evaluate the effectiveness of ML in predicting neonatal mortality risk. We compared three training approaches: a generalized model applicable across all countries, country-specific models tailored to local healthcare characteristics, and a model derived from the largest single-country dataset. Utilizing data from 2010 to 2016 for training and validation from 2017 to 2019, our analysis included 575,664 pregnancies and assessed five ML algorithms based on key neonatal health indicators recommended by the World Health Organization. Notably, the generalized model demonstrated the highest predictive performance, achieving an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.816, highlighting the benefits of leveraging a diverse dataset. Our findings advocate for the integration of generalized ML models into healthcare strategies to improve neonatal health outcomes and emphasize the importance of data diversity in reducing neonatal mortality rates.**

**Keywords** Mortality risk prediction, Artificial intelligence, Data-driven interventions, Neonatal health, Health disparities, Multicentric data

Enhancing maternal and child health is a critical objective in global health initiatives. An estimated 810 women and over 15,000 children die each day from pregnancy-related causes and preventable diseases, respectively[1]. While notable progress has been made in reducing child mortality, the Millennium Development Goal (MDG) target of a two-thirds reduction by 2015 was not fully achieved[2]. With the transition to the Sustainable Development Goals (SDGs), global efforts now focus on achieving SDG 3.2, which aims to end preventable deaths of newborns and children under five years of age. This includes reducing neonatal mortality to at least as low as 12 per 1,000 live births and under-five mortality to at least as low as 25 per 1,000 live births in all countries. Recent data indicate mixed progress, with some regions on track while others face persistent challenges in reaching these targets[3].

Artificial intelligence (AI) techniques hold significant promise for enhancing maternal and child health outcomes, particularly in low-resource environments[4]. AI encompasses a broad range of computational methods, including machine learning (ML), which allows models to automatically learn patterns from data and make predictions without being explicitly programmed[5]. Unlike traditional regression models, which rely on predefined equations and assumptions about relationships between variables, ML techniques can capture complex, nonlinear patterns and interactions within large, high-dimensional datasets. This capability is particularly valuable in maternal and child health, where multiple biological, environmental, and socioeconomic factors interact in ways that traditional statistical models may not fully capture. The increasing availability of

[1]School of Public Health, University of São Paulo, Av. Dr. Arnaldo, 715 - Cerqueira César, São Paulo 01246-904, SP, Brazil. [2]Economics Graduate Program, IDP - Brazilian Institute of Education, Development and Research, Brasilia, DF, Brazil. [3]Instituto de Pesquisa em Inteligência Artificial Aplicada à Saúde, São Paulo, Brazil. ✉email: gabriel8.silva@usp.br

maternal and child health data presents an opportunity to leverage ML-driven insights for improved healthcare decision-making, particularly in settings with limited specialized healthcare professionals.

Neonatal mortality remains a pressing public health challenge in low- and middle-income countries, driven by barriers such as restricted access to quality prenatal care, insufficient healthcare infrastructure, and a lack of specialized medical professionals[6]. Preventing neonatal mortality is crucial for improving collective health indicators and has a profound impact on families and communities. ML algorithms can analyze large volumes of data, identify patterns and risk factors, and provide valuable insights for healthcare professionals, enabling proactive and targeted interventions[7]. Deploying machine learning models offers significant potential to decrease neonatal mortality rates and enhance public health outcomes in low- and middle-income countries. In this context, multicentric studies emerge as pivotal in producing strong findings across varying socio-economic and geographical landscapes.

This study employs multicentric neonatal data from low- and middle-income countries to assess and compare various ML training strategies, aiming to identify the optimal method for neonatal mortality prediction. It specifically evaluates the efficacy of generalized models, country-specific models, and models based on the largest single-country dataset. Our goal is to improve neonatal mortality predictions, thereby advancing neonatal healthcare in settings with limited resources.

## Materials and methods

### Ethics and consent to participate declarations

All data used in this work were obtained from secondary sources. This research did not involve interaction with human participants, nor did it collect or generate any original human data. The original data[8] were collected by the Global Network for Women's and Children's Health Research (Global Network) under the supervision and guidance of ethics committees. The dataset was obtained through a formal request submitted to the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), part of the National Institutes of Health (NIH).

### Study design

This study encompassed a diverse international sample with training data drawn from the Democratic Republic of Congo, Guatemala, Zambia, India-Belagavi, India-Nagpur, Pakistan, Kenya and Argentina. We developed and assessed predictive models for neonatal death, including a general model applicable to all countries, specific models for each participating country, and a model optimized for the largest subset of training data. A synthetic flowchart is presented in Fig. 1.

### Data source

We utilized data from the Maternal Newborn Health Registry (MNHR) of the Global Network, a longitudinal, population-based study designed to track and analyze pregnancy outcomes in defined low-resource settings[9]. The study collected data from approximately 500,000 pregnancies between 2010 and 2019, in three phases. Data were collected in eight different countries: Argentina, Zambia, Guatemala, Kenya, Pakistan, India, the Democratic Republic of the Congo (DRC), and Bangladesh. The variables collected included gestational

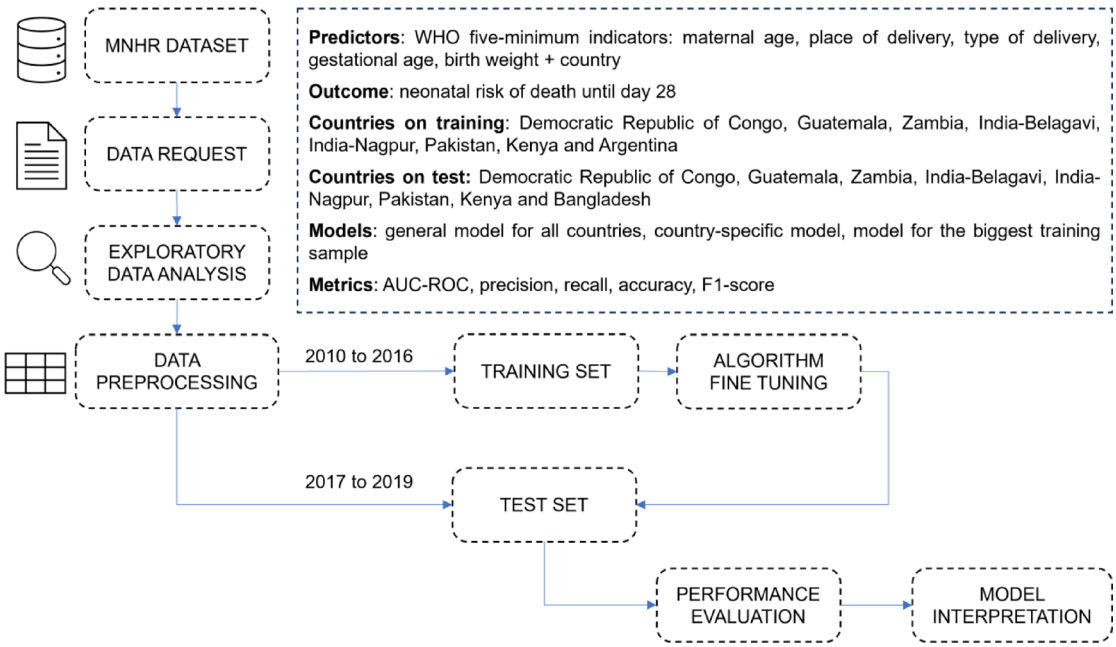


**Fig. 1**. Workflow development for neonatal risk of death prediction using machine learning algorithms and MNHR data.

information, delivery details, and a 42-day follow-up after delivery, encompassing the five minimum indicators suggested by the World Health Organization. The input data structure includes maternal age in years, place of delivery (hospital, clinic/health center, or home/other), mode of delivery (vaginal, assisted vaginal, c-section, miscarriage, medical termination of pregnancy), birth weight in grams, gestational age in weeks, and region (Argentina, DRC, Zambia, Guatemala, Bangladesh, India-Belagavi, Pakistan, India-Nagpur, Kenya).

### Machine learning techniques

In the process of developing the ML algorithm for neonatal mortality prediction, several steps were undertaken to ensure data preparation[10]. Categorical features were encoded using one-hot encoding. In order to address missing values within continuous variables, a mean imputation method was applied, wherein missing data points were replaced with the mean value of the respective features. Furthermore, for continuous features like maternal age, birth weight, and gestational age, a Z-score standardization was performed. This standardization technique normalized the values of these variables to a standard normal distribution by subtracting the mean and dividing by the standard deviation, ensuring uniformity and comparability in the dataset. To assess multicollinearity among the continuous predictors, a Variance Inflation Factor (VIF) analysis was conducted.

We tested the predictive performance of popular ML algorithms, such as Adaboost[11], XGBoost[12], CatBoost[13], LightGBM[14], Random Forest[15], and Logistic Regression. These algorithms have been carefully selected to ensure comprehensive evaluation and comparison of performance. The training strategy followed a time-period hold-out approach, where the training set was comprised of data collected from 2010 to 2016, and the test set consisted of data collected from 2017 to 2019. This approach enables the assessment of algorithm performance on different time periods, providing insights into their effectiveness in predicting neonatal risk of death across multiple years.

Hyperparameters for the algorithms were tuned using random search, which involves sampling random combinations of hyperparameters to find its optimal configuration. Model performance in the training set was assessed using a 10-fold stratified cross-validation approach, which divides the data into 10 subsets while ensuring proportional representation of different classes. This process was further repeated for 50 iterations.

Three general approaches were tested for algorithm construction: general algorithms for all countries, country-specific algorithms, and an algorithm based on the country with the largest sample in the training set. These approaches were designed to evaluate the best training strategies for the specific dataset and the generalizability capabilities of the algorithms across different countries.

Different metrics were used to evaluate the performance, robustness, and general characteristics of the AI algorithms, such as AUC-ROC (Area Under the Receiver Operating Characteristic curve), precision, recall, accuracy, and F1-score. The AUC-ROC metric is particularly significant for comparing different approaches, as it provides an overall assessment of the model's ability to distinguish between the positive and negative classes, making it a suitable criterion for decision-making. Additionally, precision measures the proportion of true positive predictions out of all positive predictions, recall evaluates the proportion of true positive predictions out of the actual positive instances, accuracy determines the overall correctness of the predictions, and the F1-score combines precision and recall assessing the model's overall performance. By employing these metrics, a comprehensive evaluation of the AI systems can be achieved, by comparing distinct countries and training strategies. In addition to evaluating performance metrics, we also assessed the calibration of the general model using calibration plots. The complete methodological process is summarized in Fig. 1. The analysis was performed using Python 3.9.21. We followed Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis – Artificial Intelligence extension (TRIPDO + AI)[16] guidelines for developing and reporting the predictive models (checklist available in the supplementary appendix – Table A.9).

### Outcome definition

The target variable of interest was neonatal status between 0 and 28 days after birth, which is categorized as either "live" or "death." As a binary feature, it consists of two distinct categories. During the training of the AI models, the focus was on predicting the risk of the neonate experiencing the specified outcome in the future. This prediction is represented as a probabilistic measure. By establishing an initial threshold, typically set at 0.5, the expected outcome for the neonate is determined based on the interaction with this measure.

### Additional models

From the overall best-performing model, we developed new models to evaluate types of variations that could further enhance performance. For this purpose, we also disaggregated the outcome temporally, considering two extra categories: death within 0–7 days from birth, i.e., within the first week of the baby's life, and death within 0–42 days from birth, in addition to the neonatal postpartum period, presented in the original model of 0–28 days.

Beyond variations by time to outcome, we tested a model that included new variables, such as ultrasound methods, the mother's educational background, prenatal visits, tetanus vaccination during pregnancy, the sex of the baby, and the trimester of pregnancy at the first prenatal visit. The purpose of including these new variables was to determine whether there is an improvement beyond the simpler model that used the five minimum indicators recommended by the WHO.

## Results

### Descriptive data analysis

After data preprocessing, the study included a total of 575,664 pregnancies. The analysis of patient distribution based on delivery location revealed that 31.3% of the births occurred in clinics or health centers, 24.2% at home or other locations, and 44.4% in hospitals. Regarding the type of delivery, 14.4% of neonates were delivered via

C-section, 84.7% through vaginal delivery, and 0.9% through vaginal-assisted delivery. Similar to the delivery location data, there were very few missing values concerning the type of delivery.

The dataset included deliveries from various countries, each contributing with different proportions of neonatal patients. Specifically, Argentina accounted for 1.7% of the cases, Bangladesh 0.2%, the Democratic Republic of Congo 6.3%, Guatemala 15.0%, India-Belagavi 21.4%, India-Nagpur 14.6%, Kenya 13.5%, Pakistan 15.8%, and Zambia 11.6%. For the training set, India-Belagavi had the largest sample size ($n = 107,076$) and was selected as a reference model to assess performance in other countries.

In terms of outcomes, 2.5% of newborns experienced neonatal death. When splitting the dataset for training and testing, the ratios were maintained similarly, at 2.6% and 2.2%, respectively. A comprehensive analysis of neonatal patient characteristics, categorized by delivery location, delivery type, country of delivery, and outcome, is presented in Table 1.

The VIF analysis was conducted on the training set to assess multicollinearity among the predictor variables. The results indicate that all variables—Maternal Age (VIF = 1.012), Gestational Age (VIF = 1.008), and Birth Weight (VIF = 1.016)—have VIF values close to 1, suggesting negligible multicollinearity. Furthermore, the correlation plot in Figure A.1 (Supplementary Appendix) confirms a low correlation between these three variables, reinforcing their independence in the dataset.

## Algorithmic performance

The performance of different ML algorithms was evaluated within each country, considering three different model approaches: General, Country-specific, and largest train size. The models were trained using lightgbm (LGBM), xgboost (XGB), adaboost, catboost and random forest, with hyperparameters optimized through random search.

In the case of Kenya (Table 2), the LGBM Tuned algorithm achieved an AUC-ROC of 0.808 [0.777, 0.839] when using the General model approach. Comparatively, the Country-specific approach with the same XGB Tuned algorithm yielded a slightly lower AUC-ROC of 0.805 [0.774, 0.836]. These results suggest that the general algorithm is the more effective approach for achieving better predictive results in Kenya. Similar trends were observed for other countries. In the Democratic Republic of the Congo (DRC), the General model (LGBM Tuned) achieved a 0.797 [0.77, 0.825] AUC-ROC and 0.793 [0.765, 0.819] for the Country-specific approach (XGB Tuned). Although the difference is marginal, the General model approach tended to outperform the Country-specific approach.

| Variable | Full Dataset | Death | Non-death | Train | Test |
|---|---|---|---|---|---|
| Place of Delivery | | | | | |
| Clinic or Health Center | 180,180 (31.3%) | 3,648 (25.6%) | 176,532 (31.4%) | 128,603 (29.1%) | 51,577 (38.5%) |
| Home or Other | 139,580 (24.2%) | 3,704 (26.0%) | 135,876 (24.2%) | 115,469 (26.1%) | 24,111 (18.0%) |
| Hospital | 255,779 (44.4%) | 6,907 (48.4%) | 248,872 (44.3%) | 197,460 (44.7%) | 58,319 (43.5%) |
| Missing value | 125 (0.0%) | 7 (0.0%) | 118 (0.0%) | 125 (0.0%) | 0 (0.0%) |
| Type of Delivery | | | | | |
| C-section | 82,900 (14.4%) | 2,262 (15.9%) | 80,638 (14.4%) | 59,631 (13.5%) | 23,269 (17.4%) |
| Vaginal | 487,631 (84.7%) | 11,708 (82.1%) | 475,923 (84.8%) | 377,215 (85.4%) | 110,416 (82.4%) |
| Vaginal Assisted | 5,108 (0.9%) | 293 (2.1%) | 4,815 (0.9%) | 4,786 (1.1%) | 322 (0.2%) |
| Missing value | 25 (0.0%) | 3 (0.0%) | 22 (0.0%) | 25 (0.0%) | 0 (0.0%) |
| Country of Delivery | | | | | |
| Argentina | 9,753 (1.7%) | 103 (0.7%) | 9,650 (1.7%) | 9,753 (2.2%) | 0 (0.0%) |
| Bangladesh | 1,089 (0.2%) | 32 (0.2%) | 1,057 (0.2%) | 0 (0.0%) | 1,089 (0.8%) |
| Democratic Republic of Congo | 36,542 (6.3%) | 884 (6.2%) | 35,658 (6.4%) | 17,974 (4.1%) | 18,568 (13.9%) |
| Guatemala | 86,298 (15.0%) | 2,050 (14.4%) | 84,248 (15.0%) | 61,131 (13.8%) | 25,167 (18.8%) |
| India-Belagavi | 123,146 (21.4%) | 2,832 (19.9%) | 120,314 (21.4%) | 107,076 (24.2%) | 16,070 (12.0%) |
| India-Nagpur | 84,046 (14.6%) | 1,760 (12.3%) | 82,286 (14.7%) | 65,768 (14.9%) | 18,278 (13.6%) |
| Kenya | 77,640 (13.5%) | 1,094 (7.7%) | 76,546 (13.6%) | 58,003 (13.1%) | 19,637 (14.7%) |
| Pakistan | 90,543 (15.7%) | 4,501 (31.6%) | 86,042 (15.3%) | 74,022 (16.8%) | 16,521 (12.3%) |
| Zambia | 66,607 (11.6%) | 1,010 (7.1%) | 65,597 (11.7%) | 47,930 (10.9%) | 18,677 (13.9%) |
| Quantitative predictors | | | | | |
| Maternal Age [mean (standard deviation)] | 24.855 (5.344) | 25.449 (5.725) | 24.838 (5.332) | 24.743 (5.218) | 25.220 (5.721) |
| Gestational Age [mean (standard deviation)] | 37.516 (5.541) | 34.339 (6.560) | 37.605 (5.481) | 37.223 (6.044) | 38.482 (3.202) |
| Birth Weight [mean (standard deviation)] | 2,911.647 (491.778) | 2,270.323 (778.100) | 2,926.885 (471.973) | 2,913.343 (493.174) | 2,906.103 (487.149) |
| Outcome | | | | | |
| Neonatal Death | 14,266 (2.5%) | - | - | 11,281 (2.6%) | 2,985 (2.2%) |
| Neonatal non-death | 561,398 (97.5%) | - | - | 430,376 (97.4%) | 131,022 (97.8%) |

**Table 1**. Descriptive summary of full, train and test datasets.

| Country | Model | Algorithm | Support | Positive Outcome | CI AUC-ROC | CI Recall | Accuracy | Precision | Specificity | F1-score |
|---|---|---|---|---|---|---|---|---|---|---|
| General[1] | General | LGBM Tuned | 134,007 | 2,985 | 0.816 [0.807, 0.825] | 0.220 [0.205, 0.235] | 0.980 | 0.642 | 0.997 | 0.328 |
| General[1] | Largest train size | LGBM | 134,007 | 2,985 | 0.76 [0.750, 0.771] | 0.998 [0.174, 0.201] | 0.980 | 0.657 | 0.188 | 0.292 |
| DRC | General | LGBM Tuned | 18,568 | 445 | 0.797 [0.770, 0.825] | 0.245 [0.207, 0.286] | 0.980 | 0.752 | 0.998 | 0.369 |
| DRC | Country-specific | XGB Tuned | 18,568 | 445 | 0.793 [0.765, 0.819] | 0.252 [0.215, 0.293] | 0.98 | 0.762 | 0.998 | 0.378 |
| DRC | Largest train size | LGBM | 18,568 | 445 | 0.749 [0.722, 0.778] | 0.997 [0.200, 0.277] | 0.979 | 0.679 | 0.238 | 0.353 |
| Guatemala | General | LGBM Tuned | 25,167 | 539 | 0.795 [0.772, 0.819] | 0.232 [0.198, 0.269] | 0.982 | 0.772 | 0.998 | 0.357 |
| Guatemala | Country-specific | LGBM Tuned | 25,167 | 539 | 0.796 [0.774, 0.82] | 0.232 [0.197, 0.269] | 0.981 | 0.706 | 0.998 | 0.349 |
| Guatemala | Largest train size | LGBM | 25,167 | 539 | 0.71 [0.681, 0.738] | 0.999 [0.123, 0.184] | 0.981 | 0.783 | 0.154 | 0.257 |
| Zambia | General | LGBM Tuned | 18,677 | 215 | 0.785 [0.745, 0.826] | 0.265 [0.206, 0.325] | 0.990 | 0.679 | 0.999 | 0.381 |
| Zambia | Country-specific | LGBM Tuned | 18,677 | 215 | 0.801 [0.761, 0.839] | 0.237 [0.181, 0.297] | 0.990 | 0.630 | 0.998 | 0.345 |
| Zambia | Largest train size | LGBM | 18,677 | 215 | 0.744 [0.701, 0.789] | 0.998 [0.176, 0.291] | 0.989 | 0.575 | 0.233 | 0.331 |
| India-Belagavi | General | LGBM Tuned | 16,070 | 313 | 0.784 [0.751, 0.814] | 0.278 [0.224, 0.328] | 0.984 | 0.725 | 0.998 | 0.402 |
| India-Belagavi | Country-specific | LGBM | 16,070 | 313 | 0.781 [0.75, 0.811] | 0.259 [0.206, 0.308] | 0.983 | 0.692 | 0.998 | 0.377 |
| India-Belagavi | Largest train size | LGBM | 16,070 | 313 | 0.781 [0.750, 0.811] | 0.998 [0.206, 0.308] | 0.983 | 0.692 | 0.259 | 0.377 |
| India-Nagpur | General | LGBM Tuned | 18,278 | 323 | 0.812 [0.781, 0.84] | 0.232 [0.190, 0.277] | 0.983 | 0.568 | 0.997 | 0.330 |
| India-Nagpur | Country-specific | XGB Tuned | 18,278 | 323 | 0.808 [0.778, 0.836] | 0.186 [0.146, 0.229] | 0.983 | 0.571 | 0.997 | 0.28 |
| India-Nagpur | Largest train size | LGBM | 18,278 | 323 | 0.804 [0.774, 0.833] | 0.996 [0.171, 0.258] | 0.982 | 0.504 | 0.214 | 0.300 |
| Pakistan | General | LGBM Tuned | 16,521 | 838 | 0.772 [0.752, 0.790] | 0.185 [0.157, 0.212] | 0.950 | 0.515 | 0.991 | 0.272 |
| Pakistan | Country-specific | LGBM Tuned | 16,521 | 838 | 0.770 [0.750, 0.788] | 0.189 [0.161, 0.215] | 0.950 | 0.506 | 0.990 | 0.275 |
| Pakistan | Largest train size | LGBM | 16,521 | 838 | 0.746 [0.726, 0.766] | 0.997 [0.122, 0.172] | 0.954 | 0.719 | 0.147 | 0.244 |
| Kenya | General | LGBM Tuned | 19,637 | 280 | 0.808 [0.777, 0.839] | 0.161 [0.119, 0.204] | 0.987 | 0.634 | 0.999 | 0.256 |
| Kenya | Country-specific | XGB Tuned | 19,637 | 280 | 0.805 [0.774, 0.836] | 0.139 [0.102, 0.179] | 0.986 | 0.557 | 0.998 | 0.223 |
| Kenya | Largest train size | LGBM | 19,637 | 280 | 0.764 [0.730, 0.797] | 0.998 [0.109, 0.190] | 0.986 | 0.583 | 0.150 | 0.239 |
| Bangladesh[2] | General | LGBM Tuned | 1,089 | 32 | 0.854 [0.779, 0.920] | 0.156 [0.038, 0.306] | 0.971 | 0.500 | 0.995 | 0.238 |
| Bangladesh[2] | Largest train size | LGBM | 1,089 | 32 | 0.793 [0.691, 0.896] | 0.999 [0.062, 0.334] | 0.975 | 0.857 | 0.188 | 0.308 |

**Table 2**. Test results for predictive models of neonatal risk of death. [1]The general model does not have a country-specific approach. [2]Due to the lack of data on training set, Bangladesh does not present a country-specific approach, since it had no data collected in the training period.

For Guatemala, the LGBM Tuned algorithm achieved AUC-ROC of 0.795 [0.772, 0.819] for the General model and a 0.796 [0.774, 0.820] AUC-ROC performance for the Country-specific model. In the case of Zambia, the General model approach with the LGBM Tuned algorithm achieved an AUC-ROC of 0.785 [0.745, 0.826], while the Country-specific approach with the XGB Tuned algorithm outperformed with an AUC-ROC of 0.801 [0.761, 0.839]. Nonetheless, we observed a better recall for the general model.

In India-Belagavi, both the General and Country-specific models, using the LGBM Tuned and LGBM algorithms, yielded comparable AUC-ROCs of 0.784 [0.751, 0.814] and 0.781 [0.75, 0.811], respectively. These findings suggest that the choice between algorithms did not significantly affect the predictive performance. Overall, the results highlight that the choice of algorithm and model approach can influence the predictive performance within each country (Fig. 2). While the General model approach generally yielded favorable results, there are instances where the Country-specific approach or a different algorithm were more effective. It is important to carefully consider the specific context and data characteristics when selecting the most suitable algorithm and model approach for each country.

In terms of calibration for the general approach (Fig. 3), we observed that the blue line, representing the LGBM model, follows the diagonal closely in the lower probability range (0–0.4), indicating good calibration. However, in the mid-range (0.4–0.7), the model slightly deviates, suggesting some overconfidence in predictions. In the higher probability range (0.7–1.0), the model gradually realigns with the diagonal, indicating improved calibration at higher confidence levels.

Figure 4 presents an analysis of the most important predictors, according to the Shapley Values, considering the general model. The variables Birth Weight and Gestational Age were the most relevant for predicting the risk of neonatal death in the general model, where the LGBM Tuned algorithm performed the best.

From Figures A.2 to A.8 (supplementary appendix), it is evident that Birth Weight was consistently the most important predictor in the Shapley values. Except for India-Belagavi, the second most relevant predictor was Gestational Age. The results from the country-specific approach support the findings of the general model.

The complete report on the performance of all tested models across different algorithms is available in the supplementary tables (A.2 to A.8).

## Model variations

We performed further analyses to assess model performance across different outcome variations, considering the general model as baseline. Figure 5 illustrates the variations in mortality within 7 and 42 days postpartum. The models exhibit considerable variability in performance across different national contexts. This variability may

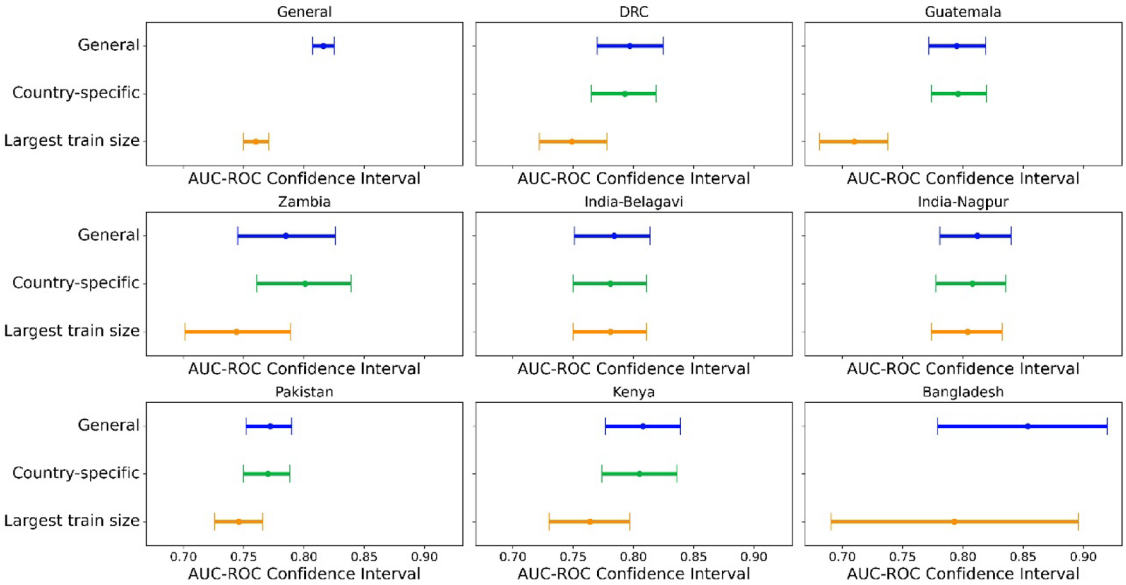


**Fig. 2**. Comparison of AUC-ROC across models and countries.

be attributed to the characteristics of the predictors in each country. Notably, the "General" model, trained with all countries together, consistently shows robust performance across the three temporal variations of outcome, suggesting its potential baseline model for neonatal mortality. However, it does not consistently outperform other models, indicating that local factors significantly influence model efficacy.

The results suggest that "Country-specific" models occasionally outperform the "General" model in certain locales, such as India-Nagpur and Bangladesh. This superior performance can likely be linked to these models' ability to accommodate localized healthcare data and demographic nuances that are not as pronounced in the global dataset. The findings advocate for the customization of predictive models to enhance their accuracy and relevance in specific regional contexts.

The stability of model performance across different time periods within the same country and model type is noteworthy. This consistency is important for the practical application of these models in healthcare settings, as it ensures reliability and predictability in their predictive capabilities over time. In general, we could see that the models exhibited higher AUC-ROC values for deaths occurring within 7 days postpartum. As the period extended, a general decline in performance was observed. The model predicting death within 42 days postpartum only outperformed others in the DRC and Bangladesh, both for the biggest training size strategy. In Guatemala, we observed the most significant discrepancy in model performances. Supplementary Table A.1 indicates that Guatemala had the lowest concentration of deaths in the 0–7 day period, accounting for approximately 59.3% of deaths during this timeframe. In the DRC, around 89% of deaths occurred within 7 days postpartum, leading to a more balanced model performance.

In addition to comparing different time periods, we also evaluated the impact of including additional variables beyond the five recommended by the WHO. This phase incorporated variables such as ultrasound methods, maternal educational background, prenatal visits, tetanus vaccination during pregnancy, infant sex, and the trimester of pregnancy at the first prenatal visit. The AUC-ROC values remained relatively similar overall.

As shown in Table 3, the most significant improvement was observed in India-Nagpur, where the AUC-ROC increased from 0.812 to 0.823. However, the DRC saw a reduction in its AUC-ROC.

## Discussion

Our study analyzed different algorithms and approaches to assess their performance in predicting the neonatal risk of death. The findings support the use of generalized ML models over country-specific or single-largest-sample models for predicting neonatal mortality. The results highlight the potential of ML in improving neonatal health outcomes by utilizing extensive and varied datasets to train predictive algorithms.

Another finding from this study was the confirmation of the importance of collecting the minimum five indicators recommended by the WHO for assessing neonatal health. These indicators provided valuable information for constructing predictive tools that can aid in clinical decision-making for the neonatal population, especially variables such as Birth Weight, Gestational Age, and Maternal Age. Additionally, factors such as the place of delivery and type of delivery were also found to be influential predictors. It is important to note that, in comparison with existing literature—including other machine learning studies on neonatal prediction[4] and meta-analyses on infant mortality evaluation[17] — similar predictive factors for neonatal outcomes were identified. Despite incorporating new variables into our models, the enhancements in predictive performance were insufficient to warrant the systematic collection of additional predictors, which would entail operational costs in clinical practice, particularly in settings with budgetary and resource constraints. Therefore, we can

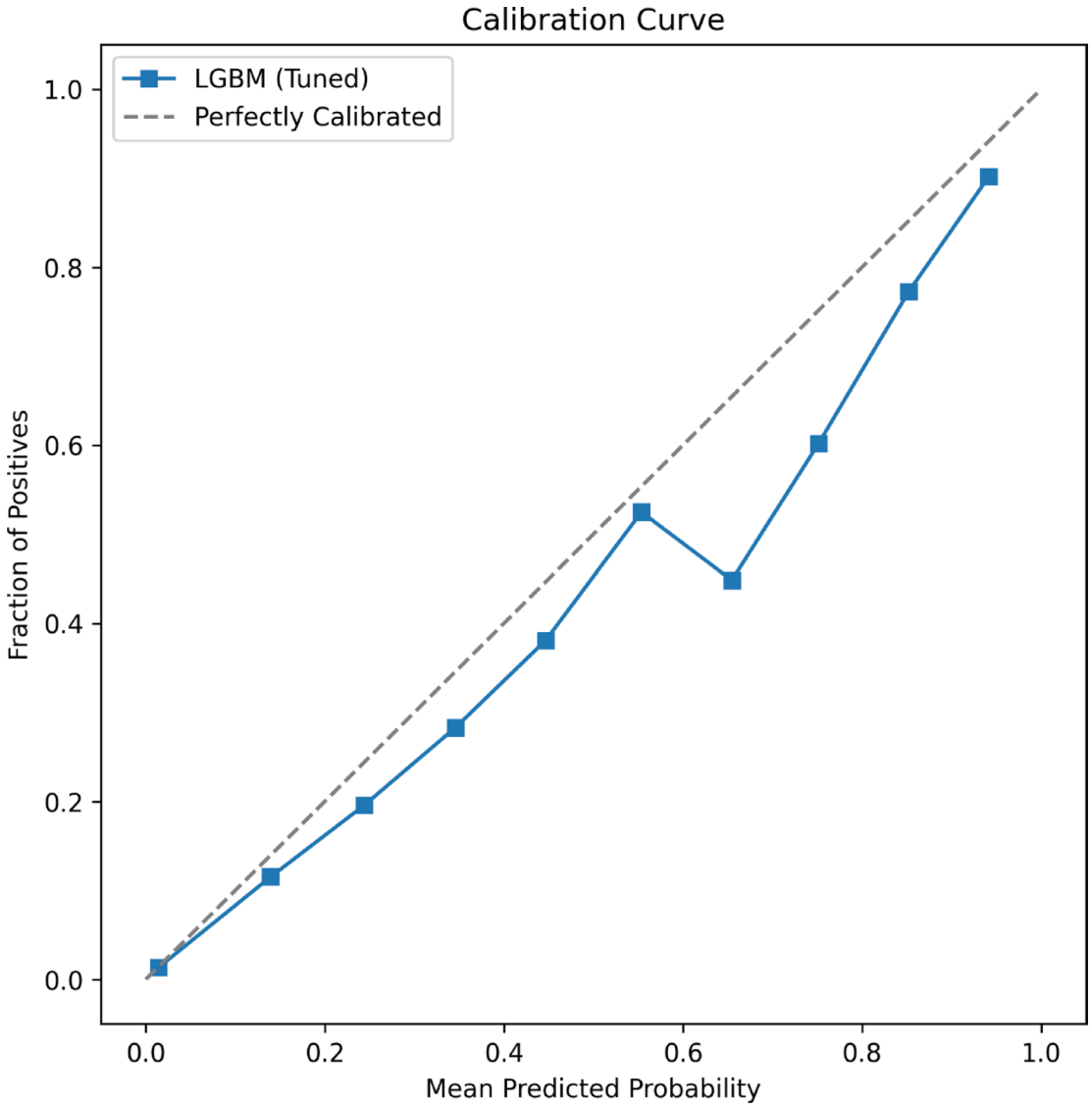


**Fig. 3**. Calibration Plot of the Tuned LightGBM Model in the general approach.

regard the five World Health Organization Minimum Indicators as essential baseline predictors for forecasting neonatal mortality.

Furthermore, by examining various temporal segments of the outcomes, we found that extending the prediction horizon to 42 days can result in a decline in the predictive performance of the models. Therefore, it may be advisable to concentrate on either the 0–7 day timeframe or the neonatal period for more accurate predictions.

Our findings align with previous studies that identified similar predictive factors for neonatal outcomes, reinforcing the value of established indicators like Birth Weight and Gestational Age[4]. A 2019 study used ML to predict postpartum hospital admission in the first 12 weeks after delivery and found a high predictive performance for hospitalization from hypertensive disorders (AUC = 0.879)[18]. Another analysis from 2020 found that ML was able to predict height-for-age z-scores in children from a rural area of Pakistan[19]. A more recent study from 2021 found that ML algorithms were able to predict with reasonable accuracy the risk of readmission for complications of hypertensive disorders of pregnancy[20]. When compared to different neonatal risk scores, this study also demonstrates good performance. Although the comparative baseline varies due to differences in populations, contexts, and variables, previous studies have reported neonatal mortality risk scores with AUC-ROC values ranging between 0.75 and 0.89[21]. However, it is important to consider that some of these scores, such as the Score for Neonatal Acute Physiology (SNAP) and SNAP Perinatal Extension (SNAPE-II), incorporate additional clinical variables, including blood pressure, temperature, oxygen saturation, and neurological reflex tests, among others.

Our study offers new insights into predicting neonatal mortality across low- and middle-income countries. Despite the promising trajectory of AI in the medical field, our research sought to explore the unique obstacles encountered by medical practitioners in resource-limited environments. Our results indicate the potential of

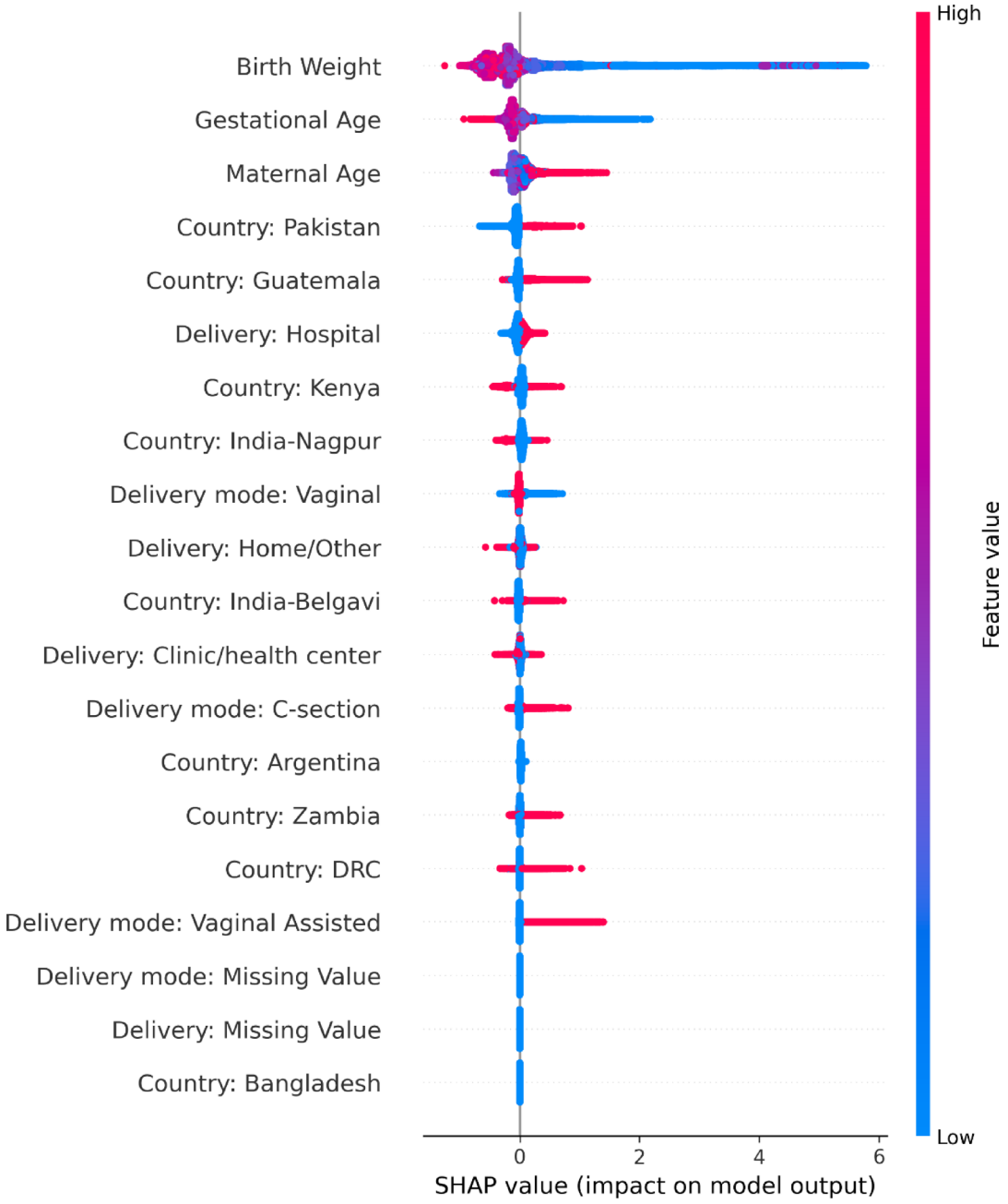


**Fig. 4**. Mean absolute Shapley-values barplot for predictors of neonatal risk of death using LGBM Tuned algorithm in the general approach.

employing machine learning to predict neonatal mortality and examines a range of training methodologies using multicenter data.

The high predictive performance of neonatal mortality models provides essential clinical and managerial insights, potentially supporting healthcare teams to take actions. For instance, it could enable the early identification of neonates at risk of mortality, allowing for the prioritization of critical interventions such as Kangaroo Mother Care[22] for preterm infants, early initiation of breastfeeding to enhance immunity, and targeted prophylactic measures, including neonatal resuscitation training for birth attendants and timely antibiotic administration for infection prevention. Neonatal mortality is primarily driven by premature birth, intrapartum-related events such as birth asphyxia, neonatal infections including sepsis and pneumonia, and congenital anomalies[23]. Implementing targeted prophylactic protocols to address these leading causes is crucial. These measures are particularly important in countries facing shortages of healthcare professionals and hospital beds. Scalable solutions, such as task-shifting strategies that train community health workers to provide essential neonatal care, telemedicine for neonatal follow-ups, and strengthening supply chains to ensure access

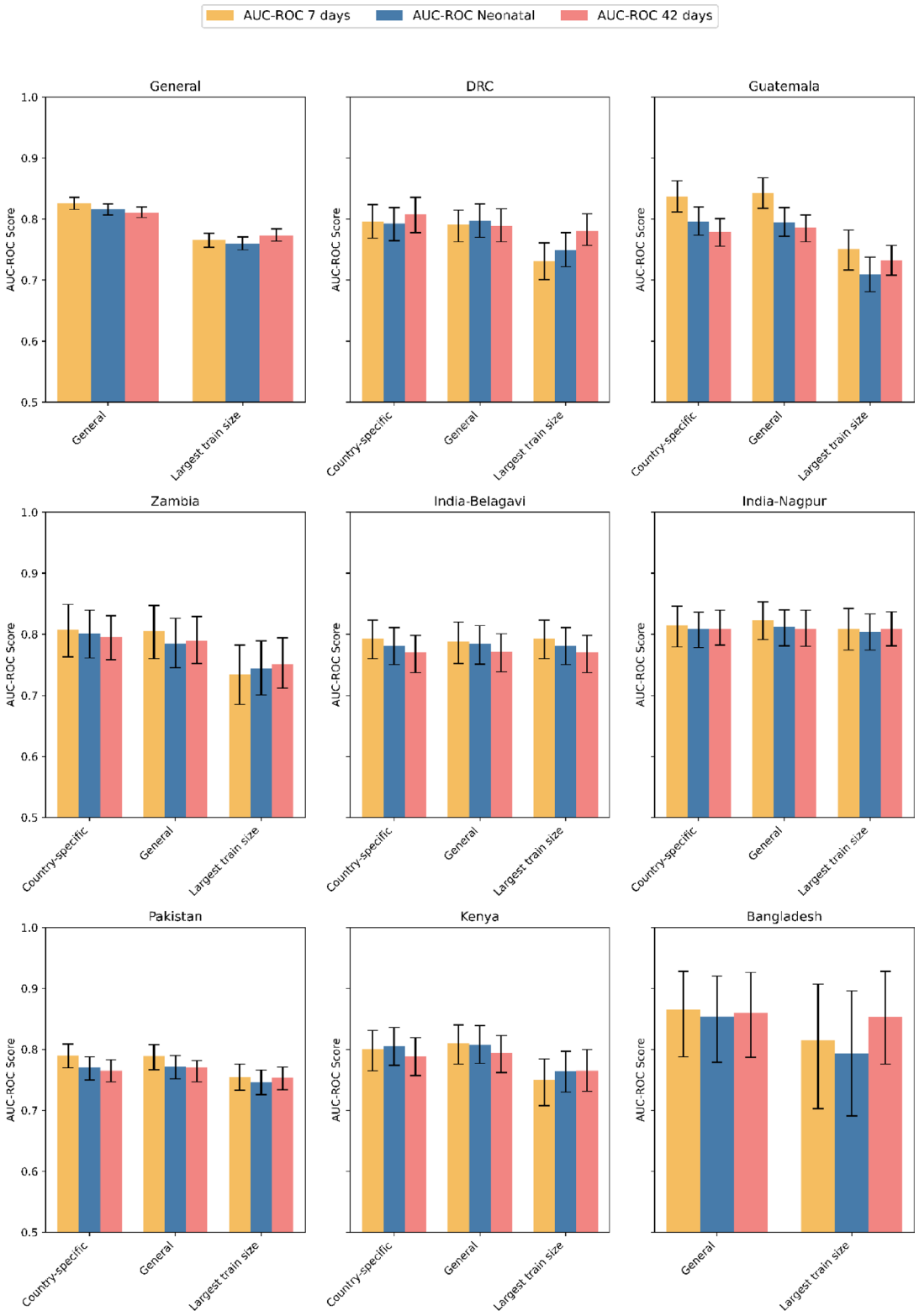


**Fig. 5**. Comparison of AUC-ROC across models, countries, and outcome variations.

to neonatal resuscitation equipment, could help bridge healthcare gaps, supported by predictive models for targeted interventions.

The findings of this study highlight the potential of ML models to aid clinical decision-making in resource-limited settings by accurately predicting neonatal mortality. By focusing on well-established indicators, healthcare providers can prioritize interventions and allocate resources more efficiently. For instance, models could support the implementation of targeted prophylactic protocols and nutritional interventions, especially

| Country | Model | Algorithm | Support | Positive Outcome | CI AUC-ROC | Comparison to neonatal baseline model |
|---|---|---|---|---|---|---|
| General[1] | General | LGBM Tuned | 134,007 | 2985 | 0.821 [0.812, 0.830] | Superior |
| General[1] | Biggest train size | Adaboost Tuned | 134,007 | 2985 | 0.788 [0.779, 0.798] | Superior |
| DRC | General | LGBM Tuned | 18,568 | 445 | 0.794 [0.767, 0.821] | Inferior |
| DRC | Country-specific | XGB Tuned | 18,568 | 445 | 0.788 [0.759, 0.816] | Inferior |
| DRC | Biggest train size | Adaboost Tuned | 18,568 | 445 | 0.782 [0.753, 0.810] | Superior |
| Guatemala | General | LGBM Tuned | 25,167 | 539 | 0.802 [0.782, 0.825] | Superior |
| Guatemala | Country-specific | XGB Tuned | 25,167 | 539 | 0.790 [0.768, 0.813] | Inferior |
| Guatemala | Biggest train size | Adaboost Tuned | 25,167 | 539 | 0.736 [0.708, 0.763] | Inferior |
| Zambia | General | LGBM Tuned | 18,677 | 215 | 0.788 [0.747, 0.828] | Superior |
| Zambia | Country-specific | XGB Tuned | 18,677 | 215 | 0.809 [0.771, 0.846] | Superior |
| Zambia | Biggest train size | Adaboost Tuned | 18,677 | 215 | 0.760 [0.716, 0.802] | Inferior |
| India-Belagavi | General | LGBM Tuned | 16,070 | 313 | 0.787 [0.756, 0.816] | Superior |
| India-Belagavi | Country-specific | Adaboost Tuned | 16,070 | 313 | 0.785 [0.754, 0.816] | Superior |
| India-Belagavi | Biggest train size | Adaboost Tuned | 16,070 | 313 | 0.785 [0.754, 0.816] | Superior |
| India-Nagpur | General | LGBM Tuned | 18,278 | 323 | 0.823 [0.791, 0.850] | Superior |
| India-Nagpur | Country-specific | Adaboost Tuned | 18,278 | 323 | 0.811 [0.78, 0.84] | Superior |
| India-Nagpur | Biggest train size | Adaboost Tuned | 18,278 | 323 | 0.819 [0.789, 0.848] | Superior |
| Pakistan | General | LGBM Tuned | 16,521 | 838 | 0.777 [0.758, 0.795] | Superior |
| Pakistan | Country-specific | LGBM Tuned | 16,521 | 838 | 0.777 [0.756, 0.794] | Superior |
| Pakistan | Biggest train size | Adaboost Tuned | 16,521 | 838 | 0.750 [0.727, 0.77] | Inferior |
| Kenya | General | LGBM Tuned | 19,637 | 280 | 0.814 [0.783, 0.844] | Superior |
| Kenya | Country-specific | LGBM Tuned | 19,637 | 280 | 0.809 [0.777, 0.838] | Superior |
| Kenya | Biggest train size | Adaboost Tuned | 19,637 | 280 | 0.777 [0.745, 0.812] | Inferior |
| Bangladesh[2] | General | LGBM Tuned | 1,089 | 32 | 0.860 [0.781, 0.926] | Superior |
| Bangladesh[2] | Biggest train size | Adaboost Tuned | 1,089 | 32 | 0.822 [0.739, 0.898] | Superior |

**Table 3**. Test results for predictive models of neonatal risk of death with the addition of new predictors. [1]The general model does not have a country-specific approach. [2]Due to the lack of data on training set, Bangladesh does not present a country-specific approach, since it had no data collected in the training period.

in environments with limited healthcare personnel and infrastructure. Although promising, the deployment of these ML models in clinical practice requires further validation through randomized trials to ensure their effectiveness and reliability in diverse clinical settings.

Future research should focus on incorporating post-pandemic data to address potential shifts in neonatal health outcomes and predictor associations. Moreover, studies should explore strategies for identifying and correcting dataset shifts, which may have implications for the long-term accuracy and applicability of predictive models. Conducting randomized clinical trials to evaluate the real-world impact of AI-based decision support systems on neonatal outcomes is also recommended to bridge the gap between research and clinical application. Additionally, an important avenue for future research is predicting the survival time of neonates with a predicted mortality outcome, providing further insights for improving neonatal care and clinical decision-making.

The study's strengths include the use of comprehensive, multicentric datasets and a focus on established predictive indicators. Despite the advancements and insights provided by our study, certain limitations should be acknowledged. Firstly, we did not evaluate the representativeness of the samples relative to the entire population of the included countries. As a result, we cannot claim that the findings are broadly applicable across these nations. Additionally, since the data extended only through 2019, changes in outcome distribution and the associations among predictor variables could have occurred in subsequent years, influenced by the COVID-19 pandemic and initiatives aimed at reducing neonatal mortality. These limitations suggest that while the findings are valuable, they should be interpreted with caution, particularly in the context of current clinical conditions.

## Conclusions

The study provided valuable insights into predictive modeling for neonatal mortality risk. The findings indicate the importance of specific predictive factors and the strategic selection of algorithms and training data, highlighting the value of leveraging comprehensive and multicentric data for improving predictive performance. These results can guide the development of more accurate and effective predictive models to enhance clinical decision-making, ultimately improving outcomes for neonatal populations.

## Data availability

The data used in this research is sourced from the Maternal and Neonatal Health Registry (MNHR) at the National Institutes of Health (NIH). Access to the dataset can be requested through official data request procedures established by the institute. As authors, we are not permitted to share the data publicly. The access to the code used in the development of the models can be obtained upon request to Gabriel Silva (gabriel8.silva@usp.br).

Received: 2 November 2024; Accepted: 23 May 2025
Published online: 07 July 2025

## References

1. World Health Organization, WHO. Maternal mortality: Evidence brief. Retrieved from: https://iris.who.int/bitstream/handle/10665/329886/WHO-RHR-19.20-eng.pdf (2019).
2. World Health Organization, WHO. Millennium Development Goals (MDGs). Retrieved from: https://www.who.int/news-room/fact-sheets/detail/millennium-development-goals-(mdgs) (2018).
3. UNICEF, & World Health Organization. Levels and Trends Child Mortality-Report 2023: Estimates Developed by the United Nations Inter-Agency Group for Child Mortality Estimation. (2024).
4. Batista, A. F., Diniz, C. S., Bonilha, E. A., Kawachi, I. & Chiavegatto Filho, A. D. Neonatal mortality prediction with routinely collected data: a machine learning approach. *BMC Pediatr.* **21** (1), 1–6. https://doi.org/10.1186/s12887-021-02788-9 (2021).
5. Sarker, I. H. Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.* **2** (3), 160 (2021).
6. Rosa-Mangeret, F. et al. 2.5 Million annual deaths—are neonates in low-and middle-income countries too small to be seen? A bottom-up overview on neonatal Morbi-mortality. *Trop. Med. Infect. Disease*. **7** (5), 64. https://doi.org/10.3390/tropicalmed7050064 (2022).
7. Ibrahim, M. S. & Saber, S. Machine learning and predictive analytics: advancing disease prevention in healthcare. *J. Contemp. Healthc. Analytics*. **7** (1), 53–71 (2023).
8. Koso-Thomas, M. & Nolen, T. Global Network's Maternal Newborn Health Registry (Version 1) [dataset]. NICHD Data and Specimen Hub. https://doi.org/10.57982/t880-rf36 (2019).
9. McClure, E. M. et al. The global network maternal newborn health registry: a multi-country, community-based registry of pregnancy outcomes. *Reproductive Health*. **17**, 1–11. https://doi.org/10.1186/s12978-020-01020-8 (2020).
10. Vokinger, K. N., Feuerriegel, S. & Kesselheim, A. S. Mitigating bias in machine learning for medicine. *Commun. Med.* **1** (1), 25. https://doi.org/10.1038/s43856-021-00028-w (2021).
11. Schapire, R. E. Explaining adaboost. In Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik (37–52). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-41136-6_5 (2013).
12. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In: P*roceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794) https://doi.org/10.1145/2939672.2939785 August (2016).
13. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. *Adv. Neural. Inf. Process. Syst.* **31**. https://doi.org/10.48550/arXiv.1706.09516 (2018).
14. Ke, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural. Inf. Process. Syst.* **30**. https://doi.org/10.5555/3294996.3295074 (2017).
15. Ho, T. K. Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition* Vol. 1, pp. 278–282. https://doi.org/10.1109/ICDAR.1995.598994 (1995).
16. Collins, G. S., et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. **385**, e078378 (2024).
17. Garcia, L. P., Fernandes, C. M. & Traebert, J. Risk factors for neonatal death in the capital City with the lowest infant mortality rate in Brazil. *Jornal De Pediatria*. **95**, 194–200. https://doi.org/10.1016/j.jped.2017.12.007 (2019).
18. Betts, K. S., Kisely, S. & Alati, R. Predicting common maternal postpartum complications: leveraging health administrative data and machine learning. *BJOG: Int. J. Obstet. Gynecol.* **126** (6), 702–709. https://doi.org/10.1111/1471-0528.15607 (2019).
19. Harrison, E., et al. Machine learning model demonstrates stunting at birth and systemic inflammatory biomarkers as predictors of subsequent infant growth–a four-year prospective study. *BMC pediatrics*, **20**, 1–10. https://doi.org/10.1186/s12887-020-02392-3 (2020).
20. Hoffman, M. K., Ma, N. & Roberts, A. A machine learning algorithm for predicting maternal readmission for hypertensive disorders of pregnancy. *Am. J. Obstet. Gynecol. MFM*. **3** (1), 100250. https://doi.org/10.1016/j.ajogmf.2020.100250 (2021).
21. Veloso, F. C., Barros, C. R., Kassar, S. B. & Gurgel, R. Q. Neonatal death prediction scores: a systematic review and meta-analysis. *BMJ Paediatrics Open.*, **8**(1), e003067. (2024).
22. Sivanandan, S. & Sankar, M. J. Kangaroo mother care for preterm or low birth weight infants: a systematic review and meta-analysis. *BMJ Global Health*, **8**(6), e010728. (2023).
23. World Health Organization, WHO. Newborn mortality. Retrieved from: https://www.who.int/news-room/fact-sheets/detail/newborn-mortality#:~:text=Premature%20birth%2C%20birth%20complications%20(birth,leading%20causes%20of%20neonatal%20deaths (2024).

## Acknowledgements

Funding for this research was provided by the National Council for Scientific and Technological Development – CNPq and the Department of Science and Technology of Secretariat of Science, Technology, Innovation and Health Complex of Ministry of Health of Brazil – MoH, under grant No. 445020/2023-7. This work was supported by the Instituto de Pesquisa em Inteligência Artificial Aplicada à Saúde, São Paulo, Brazil (http://www.iasaude.org.br). We express our sincere gratitude for their financial assistance, which made this research possible. We would also like to thank all the researchers and staff members who contributed their expertise and support throughout the study. The authors thank all the communities, hospitals, health providers, research staff, women and their babies who participated in the Global Network for Women's and Children's Health Research Maternal and Neonatal Registry. The study that generated the dataset used for the analyses was funded by grants from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. We acknowledge NICHD DASH for providing the Global Network's Maternal Newborn Health Registry data that was used for this research.

## Author contributions

Conceptualization: G.F.S.S., A.D.P.C-F.; methodology: G.F.S.S., A.D.P.C-F; software: G.F.S.S; validation: R.M.W., F.C.S.J., A.D.P.C-F; formal analysis: G.F.S.S.; investigation: G.F.S.S; resources: R.M.W., F.C.S.J., A.D.P.C-F; data curation: G.F.S.S; writing—original draft preparation: G.F.S.S; writing—review and editing: R.M.W., F.C.S.J., A.D.P.C-F; visualization: G.F.S.S.; supervision: R.M.W., F.C.S.J., A.D.P.C-F; project administration: R.M.W., F.C.S.J., A.D.P.C-F; funding acquisition: R.M.W., F.C.S.J., A.D.P.C-F.

## Declarations

### Competing interests
The authors declare no competing interests.

### Additional information
**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-04066-5.

**Correspondence** and requests for materials should be addressed to G.F.d.S.S.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.