# Predicting Bronchopulmonary Dysplasia in Infants: A Comparative Evaluation of Probit and Machine Learning Models

**S U Madaki[1,*], A B Muhammad[2], and H A Hamisu[1]**

[1]Department of Mathematics and Statistics, Kaduna Polytechnic, Nigeria
[2]Department of Statistics, Aliko Dangote University of Science and Technology, Nigeria

*Corresponding author's email: shazaliumar6@gmail.com

**Abstract.** This study compares the predictive performance of traditional Probit regression and several machine learning models in predicting Bronchopulmonary Dysplasia (BPD) among preterm infants. The models were evaluated using standard performance metrics, including accuracy, precision, specificity, sensitivity, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Among all models, the Random Forest demonstrated superior predictive performance with the highest accuracy (86.36%), precision (85.71%), specificity (87.50%), sensitivity (85.71%), F1-score (0.8571), and AUC (0.92), indicating a strong discriminative ability. Birth weight and postnatal weight at four weeks emerged as the most significant predictors of BPD. The findings suggest that machine learning approaches, particularly the Random Forest algorithm, provide a more robust predictive framework than the conventional Probit regression model for early detection of BPD risk in preterm infants.
**Keyword:** Birth Weight, Bronchopulmonary Dysplasia, Machine Learning, Predictive Modelling, Probit Regression

## 1. Introduction

Bronchopulmonary dysplasia (BPD) is a chronic lung disease and one of the most prevalent morbidities among preterm neonates, especially those born before 33 weeks of gestation. Despite significant advances in neonatal intensive care, BPD continues to be a major cause of neonatal morbidity and mortality worldwide [6], [9], [12], [13]. As one of the most common and clinically significant sequelae of prematurity, BPD poses substantial challenges to neonatal clinicians and researchers. Accurate identification of infants at high risk of developing BPD would facilitate timely intervention and individualised care, thus improving survival and long-term outcomes.

Most currently available BPD prediction models employ traditional multivariable logistic regression techniques. However, these models are often limited in their adaptability and generalizability. They typically fail to accommodate changing risk profiles across different gestational ages or diverse clinical populations, even though early-life exposures such as maternal health, antenatal corticosteroid use, and early respiratory support have a significant influence on BPD risk [1], [2]. Additionally, most logistic regression-based models assume linear relationships and independence among predictors, which may not adequately capture the complex interactions among clinical variables influencing BPD development [3].

Recent research has introduced machine learning (ML) techniques to overcome these limitations. For instance, [4] compared multivariable logistic regression and several machine learning algorithms for predicting BPD in preterm infants. Their comparison, based on the area under the curve (AUC), demonstrated that ML algorithms slightly outperformed logistic regression. However, the study's reliance solely on AUC limited its comprehensiveness, as it did not evaluate key diagnostic parameters such as sensitivity, specificity, F1-score, or precision metrics that are critical for assessing model robustness in clinical decision making.

Similarly, [5] utilised perinatal and postnatal factors to develop an ensemble logistic regression model for predicting BPD. Their approach combined separate perinatal and early respiratory models, validated through simulated clinical trials, achieving AUCs of 0.921 and 0.899 in training and validation datasets, respectively. However, this approach also lacked external validation, raising concerns about its applicability across diverse populations. [6], in a large multicentre cohort, they found that their day-one clinical model achieved a C-statistic of approximately 0.76, though its predictive accuracy diminished when applied to newer cohorts, indicating poor calibration over time.

In a recent systematic review, [7] reported that over 90% of published BPD prediction models employed logistic regression, with few adopting advanced machine learning or deep learning approaches. Moreover, most studies have demonstrated a high risk of bias and insufficient external validation, which limits their clinical translation. To address these shortcomings, [8] applied machine learning techniques to a high-altitude neonatal cohort, emphasizing the value of interpretable ML models tailored to local settings. Their findings suggest that machine learning can enhance the prediction of BPD, particularly when integrating environmental and regional factors.

Furthermore, several recent works have incorporated deep learning into neonatal respiratory outcome prediction. [9] Applied a convolutional neural network (DenseNet121) to chest radiographs for early BPD detection, achieving AUCs between 0.79 and 0.97 across different time points. Similarly, developed a two-stage multilayer perceptron (MLP) model based on respiratory support duration, reporting AUROCs up to 0.897 for predicting both occurrence and severity of BPD in very-low-birth-weight infants. These findings underscore the potential of ML and deep learning methods to capture nonlinear, high-dimensional interactions that traditional regression models might overlook.

In addition to these clinical approaches, emerging genomic and transcriptomic studies have also contributed to understanding the molecular basis of BPD. For instance, [11] employed competing endogenous RNA (ceRNA) co-expression networks to identify differentially expressed miRNAs associated with BPD. Gene Ontology (GO) and KEGG analyses revealed several regulatory pathways linked to inflammation and oxidative stress, offering molecular insights into BPD pathogenesis.

Building upon this literature, the present study aims to compare the predictive performance of traditional probit regression and machine learning algorithms in forecasting BPD outcomes. Unlike previous studies that relied solely on AUC, this research adopts a comprehensive evaluation framework using sensitivity, specificity, accuracy, precision, F1-score, and AUC-ROC. The overarching objective is to determine whether machine learning models can outperform classical regression in predicting BPD and to identify the most influential predictors of the disease. Ultimately, these findings are expected to enhance predictive modelling and guide early clinical intervention for at-risk neonates.

## 2. Research Methods

### *2.1. Study Design and Data Description*

This study employed a comparative analytical design using secondary data obtained from the Neonatal Intensive Care Unit (NICU) of Abdullahi Wase Specialist Hospital, Kano, Nigeria. The dataset comprised clinical records of preterm infants with variables including birth weight (grams), weight after four weeks (grams), and gender. The binary outcome variable indicated the presence (1) or absence (0) of BPD. All analyses were conducted using a 70:30 training–testing split to ensure unbiased model evaluation.

*2.2. Analytical Framework*

Two major modelling frameworks were adopted in this study: the traditional Probit regression model and a set of machine learning algorithms, including Random Forest, Logistic Regression, Support Vector Machine, and Decision Tree. The Probit regression model operates under the assumption of normally distributed errors and is used to estimate the probability that an infant develops bronchopulmonary dysplasia (BPD) based on a given set of predictor variables. In contrast, the machine learning models represent data-driven, non-parametric approaches that are capable of capturing complex and nonlinear relationships among predictors.

*2.2.1. Probit Regression*

The probit is an alternative to the logit method, but they differ in assuming a normal distribution of the random variable. The differences lie in the fact that the logistic function is a harder part, but there are no significant differences in practice, only in the case that the sample contains numerous observations with extreme values. The comparison of parameters between them cannot be directly compared because the Logistic distribution has equal variance $\pi^2/3$. In addition, the estimate attained by logit would be multiplied by $\pi^2/3$ in order to be comparable with estimates obtained in the probit model [11].

The assumptions of probit regression are: The outcome is binary, the probit of the outcome and independent variable have a linear relationship, normally distributed errors are independent, and no severe multicollinearity. These are the assumptions of traditional probit regression, while machine learning models do not depend on assumptions.

*2.2.2. Logistic Regression*

The logistic regression is used for classification tasks, and its goal is to predict the likelihood that a case belongs to a given class or not. Logistic regression is a supervised machine learning used for binary classification where its sigmoid function used takes input as independent factors and to produce a probability value between 0 and 1 [5].

The equation of logistic regression will be:

$$P(X,b,w) = \frac{e^{w.X+b}}{1+e^{w.X+b}} = \frac{1}{1+e^{-w.X+b}} \quad (1)$$

*2.2.3. Random forest*

Random forest is a predictor consisting of a combination of randomised base regression trees $\{r_k(\mathrm{x}\Theta_m, \mathcal{D}_k), m \geq 1\}$, where $\theta_1, \theta_2 \ldots$ i.id. Are outcomes of randomised variable$\Theta$. These random trees are joined to form an aggregated regression estimate [3]

$$\bar{r_k}(\mathrm{X}, \mathcal{D}_k) = \mathbb{E}_\Theta[r_k(\mathrm{X}, \Theta, \mathcal{D}_k)] \quad (2)$$

where $\mathbb{E}_\Theta$ denotes expectation with respect to the random parameter, conditionally on X and the data set $\mathcal{D}_k$.

*2.2.4. Decision tree*

The decision tree is among the most powerful and popular classifiers used for handling problems because of its numerous benefits, which include its simple architecture, high performance and adaptability. The decision is frequently applied to overcome the classification problems in the field of data mining and machine learning [3].

*2.2.5 Support vector machine*

The support vector machine is also a supervised learning machine known as SVMs, which are used for classification as well as regression purposes. And are the data point which lie close to the hyper plane. In addition, if a data fed then the algorithm builds a classifier which can be utilize to assign new examples to one class or another [3].

*2.2.6. Model Performance Evaluation Metrics*

Model performance was assessed using the following metrics: Accuracy, Precision, Sensitivity, Specificity, F1-score, and AUC-ROC. These metrics provide a comprehensive evaluation of the models' discriminative and predictive capabilities.

Sensitivity, also known as the True Positive Rate, refers to the proportion of actual positive cases that are correctly identified by the model. It is calculated using the following formula:

$$sensitivity = \frac{TP}{TN+FN} \times 100 \quad (3)$$

Specificity (True Negative Rate) is the proportion of actual negatives correctly identified by the model, defined as

$$specificity = \frac{TN}{TN+FP} \times 100 \quad (4)$$

Precision is the proportion of true positives among all predicted positive cases. It measures the model ability to avoid false positives. The formula is

$$Precision = \frac{TP}{(TP+FP)} \quad (5)$$

Accuracy is defined as the overall proportion of correctly classified cases (both true positives and true negatives) out of cases. It measures the overall correctness of the model. The formula is

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (6)$$

The F1-score represents the harmonic mean of sensitivity and specificity, providing a balanced measure that reflects both the model's ability to correctly identify positive cases and its precision in avoiding false positives. It is calculated using the following formula:

$$F1 - Score = \frac{(SE \times SP)}{(SE+SP)} \times 100 \quad (7)$$

AUC (Area under curve) refers to the area under the receiver operating characteristic (ROC) curve. It measures a model ability to distinguish between positive and negative cases.

## 3. Result and Discussion

This section presents a comprehensive comparison of the predictive performance of probit regression and several machine learning models in forecasting the occurrence of bronchopulmonary dysplasia (BPD) among preterm infants. The analysis includes a detailed evaluation of each model's diagnostic accuracy using multiple statistical metrics, followed by an in-depth examination of variable importance to identify the most influential predictors contributing to the risk of developing BPD.

**Table 1.** Model Evaluation Results (70%Training, 30%Testing)

| Model | Accuracy | Precision | Specificity | Sensitivity | F1 Score | AUC |
|---|---|---|---|---|---|---|
| Probit Regression | 0.8182 | 0.8333 | 0.8333 | 0.8333 | 0.8333 | 0.8890 |
| Logistic Regression | 0.8182 | 0.8333 | 0.8333 | 0.8333 | 0.8333 | 0.8890 |
| Decision Tree | 0.7727 | 0.7500 | 0.7500 | 0.7500 | 0.7500 | 0.7800 |
| Random Forest | 0.8636 | 0.8571 | 0.8750 | 0.8571 | 0.8571 | 0.9200 |
| Linear SVM | 0.8182 | 0.8333 | 0.8333 | 0.8333 | 0.8333 | 0.8890 |

The table 1 above shows a comprehensive comparison of the traditional probit regression and machine learning model. The random forest appeared as the best model in terms of predictive accuracy of (0.8636), specificity (0.8750), sensitivity (0.8571), precision (0.8571), F1-score (0.9200) and Auc of (0.9200) which indicate excellent in discriminating ability. This high performance stems from its ensemble structure, where multiple decision trees collectively reduce variance and prevent overfitting. Each tree independently contributes to the final decision, improving model robustness and generalization. The Random Forest also captures nonlinear and complex feature interactions more effectively than single estimators, which is crucial when modeling heterogeneous biomedical data such as clinical variables from preterm infants. Similar outcomes have been reported in prior research, where ensemble methods consistently outperformed individual classifiers due to their ability to represent multifactorial relationships [10], [4].

All others models in traditional and machine learning have shown promise in predictive accuracy and discriminating ability, the probit, logistic and SVM models have shown similar performance with the same performance and identical score in all metrics. This convergence suggests that the dataset's predictors may be linearly separable, with limited nonlinear interactions. Both Probit and Logistic Regression rely on similar statistical assumptions binary outcome, linearity in the link function, and normally distributed errors while a linear-kernel SVM behaves analogously under such conditions. The uniformity of their results indicates that nonlinear extensions offered no substantial advantage, implying that a linear boundary can effectively represent the relationship between the explanatory variables and BPD risk.

While the Decision Tree model show relatively weaker predictive power. Its reduced performance is mainly attributed to overfitting and sensitivity to small data fluctuations, characteristics typical of single-tree classifiers. Decision Trees recursively partition data to create homogeneous subgroups but often model noise instead of meaningful structure, yielding high variance and poor generalization on unseen data. Consequently, although Decision Trees are intuitively interpretable and computationally efficient, their predictive accuracy remains inferior to ensemble techniques like Random Forest, which effectively stabilize predictions by averaging multiple weak learners and thus minimize model variance.

**Table 2.** Relative importance of Variables.

| **Feature** | Probit | Logistic | Decision Tree | Random Forest | Linear SVM |
|---|---|---|---|---|---|
| **Gender_(M)** | 0.2100 | 0.1950 | 0.0000 | 0.0250 | 0.2050 |
| **Weight at birth** | 0.3900 | 0.4050 | 0.4700 | 0.4400 | 0.3950 |
| **Weight after four weeks** | 0.4000 | 0.4000 | 0.5300 | 0.5350 | 0.4000 |

The findings in table 2 reveal that both birth weight and weight after four weeks were the most significant predictors of Bronchopulmonary Dysplasia (BPD) across all the evaluated models, while gender exhibited only a minor influence on prediction outcomes. From a clinical perspective, this observation aligns with established neonatal research indicating that postnatal weight gain serves as a critical marker of lung maturity and overall physiological resilience in preterm infants. Infants who exhibit suboptimal weight gain during the first month of life often experience prolonged respiratory support and oxygen dependency, which increases the likelihood of developing BPD. Consequently, weight after four weeks may provide a more dynamic and integrative measure of neonatal health than birth weight, as it reflects both initial vulnerability and subsequent adaptation to extrauterine life. In contrast, birth weight captures only the intrauterine growth condition and may not fully account for postnatal complications or recovery trajectories [6],[7].

## 4. Conclusions

The analysis shows that Random Forest, which is a machine learning model, outperformed probit regression and other machine learning models by achieving the highest predictive accuracy and performance score. In contrast, probit regression, logistic regression and SVM model show similar performance, while the decision tree model performs relatively low compared to them. Notably, weight features are the most significant predictors, while gender is relatively less important in predicting BPD. Overall, the machine learning models outperformed traditional probit and other models.

## References


[1] C Siffel, K D Kistler, J F Lewis and S P Sarda "Global incidence of bronchopulmonary dysplasia among extremely preterm infants: a systematic literature review. *The journal of material fetal and neonatal medicine*, ***34(11)***:*1721-1731,2021*, doi: *https://doi.org/*10.1080/14767058.2019.1646240

[2] F Aaboub, H Chamla, and T ouaderhman,"Statistical analysis of various splitting criteria for decision trees*", journal of Algorithms & Computational Technology,* ***vol.17****, Oct.2023.[online]. Available: https://doi.org/10.1177/17483026231198181*

[3] G Biau," Analysis of a random forests model," journal of machine learning Research, vol.13, pp.1063-1095, 2012.

[4] K Faiza, C Heelen, K Amal, M Jonathan, YU Joseph, Ting, W Jonathan, and S S Prakash "Comparison of Multivariable Logistic Regression and Machine Learning Models for Predicting Bronchopulmonary Dysplasia or Death in Very Preterm Infants", journal of Font Pediatr. 2021. doi: https://doi.org/10.3389/fped.2021.759776

[5] L Jiu, J Wang, S, F J Somolinos, J Tapia-Galisteo, G Garcia-Saez, M Hernando, X Li, R A Vreman, , A K Mantel-Teeuwisse, and W G Goettsch, "A literature review of quality assessment and applicability to hta of risk prediction models of coronary heart disease in patients with diabetes," Diabetes research and clinical practice, 209:111574, 2024

[6] L M Davidson, & S K Berkelhamer "Bronchopulmonary dysplasia: chronic lung disease of infancy and long-term pulmonary outcomes". Journal of clinical medicine, 6(1)4 2017.

[7] M D Tao, H Xian, and G Wan-liang "Predictors of Bronchopulmonary Dysplasia in 625 Neonates with Respiratory Distress Syndrome" Journal of Tropical Pediatrics, 68(3), 1–10, 2022. https://doi.org/10.1093/tropej/fmac037

[8] P Kumar, P S Ambeker, M Kumar, and S Roy, "Analytical Statistics Techniques of Classification and Regression in Machine learning," in Data Mining, inTechOpen, , doi: 10.5772/intechopen 84922, 2020.

[9] Q Zhou, H B Kong, B.-M, S Y Zhou,"Biblometric analysis of bronchopulmonary dysplasia in extremely premature infants in the Web of Science database using Citespace Software, "Font. Pediatr., vol. 9, p.705033,2021

[10] R M Leigh, P Andrew, S R Srinandini, M V Farha, H Hou, K Chelsea, R Abigail, N Arvind, B C John, C Fu-sheng," Machine learning for prediction of bronchopulmonary dysplasia- free survival among very preterm infants, "BMC Paediatric., vol. 22, no. 542, pp. 1-13, 2022.

[11] T Kliestik, K K Ocisova, and M Misankova, "Logit and Probit Model used for Prediction of Financial Health of Company," Procedia Economics and Finance, vol. 23, pp. 850-855, 2015, doi:1016/S2212-5671(15)00485-2.

[12] T Sun, Yu, H.-Y, M Yang, Y F Song, and Fu, J.-H. "Risk of asthma in preterm infants with bronchopulmonary dysplasia: a systematic review and meta-analysis". World Journal of Paediatrics, 19(6):549-556, 2023.

[13] Y Zhang, J Zhang, and X Wang "Functional analysis of bronchopulmonary dysplasia- related neuropeptides in preterm infants and miRNA-based diagnostic model construction". Computational and mathematical methods in medicine, 2022.