Original Article
Yonsei Med J 2022 Jul;63(7):640-647
https://doi.org/10.3349/ymj.2022.63.7.640
Yonsei Medical Journal
YMJ
pISSN: 0513-5796 · eISSN: 1976-2437

Application of Machine Learning Approaches to Predict Postnatal Growth Failure in Very Low Birth Weight Infants

Jung Ho Han[1], So Jin Yoon[1], Hye Sun Lee[2], Goeun Park[2], Joohee Lim[1], Jeong Eun Shin[1], Ho Seon Eun[1], Min Soo Park[1], and Soon Min Lee[1]

[1]Department of Pediatrics, Yonsei University College of Medicine, Seoul;
[2]Biostatistics Collaboration Unit, Yonsei University College of Medicine, Seoul, Korea.

Purpose: The aims of the study were to develop and evaluate a machine learning model with which to predict postnatal growth failure (PGF) among very low birth weight (VLBW) infants.
Materials and Methods: Of 10425 VLBW infants registered in the Korean Neonatal Network between 2013 and 2017, 7954 infants were included. PGF was defined as a decrease in Z score >1.28 at discharge, compared to that at birth. Six metrics [area under the receiver operating characteristic curve (AUROC), accuracy, precision, sensitivity, specificity, and F1 score] were obtained at five time points (at birth, 7 days, 14 days, 28 days after birth, and at discharge). Machine learning models were built using four different techniques [extreme gradient boosting (XGB), random forest, support vector machine, and convolutional neural network] to compare against the conventional multiple logistic regression (MLR) model.
Results: The XGB algorithm showed the best performance with all six metrics across the board. When compared with MLR, XGB showed a significantly higher AUROC (p=0.03) for Day 7, which was the primary performance metric. Using optimal cut-off points, for Day 7, XGB still showed better performances in terms of AUROC (0.74), accuracy (0.68), and F1 score (0.67). AUROC values seemed to increase slightly from birth to 7 days after birth with significance, almost reaching a plateau after 7 days after birth.
Conclusion: We have shown the possibility of predicting PGF through machine learning algorithms, especially XGB. Such models may help neonatologists in the early diagnosis of high-risk infants for PGF for early intervention.

Key Words: Growth failure, very low birth weight infants, machine learning, prediction, neonatal intensive care unit

INTRODUCTION

Owing to improvements in the survival rates of preterm infants, there has been an increasing focus on their growth and neurocognitive development. Based on multiple population-based studies, a significant improvement in growth in hospitalized very low birth weight (VLBW) infants has been achieved;[1,2] however, this are still issues. In a multicenter study using the database of the Vermont Oxford Network consisting of data of VLBW infants, half of them showed postnatal growth failure (PGF) in 2013.[1] In South Korea, the overall incidence of PGF, defined as a decrease in weight Z score between birth and discharge of more than -1.28 using the Fenton growth chart, was noted as 45.5%, based on the Korean Neonatal Network (KNN) database from 2013 to 2014.[2]

Growth assessment is necessary to elucidate the extent to which an infant's nutritional needs are being met and to identify infants with difficulties in overcoming neonatal morbidities. Differences in postnatal growth rates have been shown to be associated with sex, nutritional factors, and common preterm morbidities, such as chronic lung disease, necrotizing en-

Received: December 24, 2021 **Revised:** March 22, 2022
Accepted: March 25, 2022
Corresponding author: Soon Min Lee, MD, PhD, Department of Pediatrics, Yonsei University College of Medicine, 211 Eonju-ro Gangnam-gu, Seoul 06273, Korea. Tel: 82-2-2019-3350, Fax: 82-2-2019-4881, E-mail: smlee@yuhs.ac

The results of this study were orally presented at the 34th Annual Autumn Meeting of the Korean Society of Perinatology.
•The authors have no potential conflicts of interest to disclose.

terocolitis, and sepsis.[2-6] Although we cannot completely explain the "cause and effect" relationship between preterm morbidities and PGF, there are sound physiological reasons why some comorbidities might be the cause of poor postnatal growth. Most importantly, PGF eventually adversely affects long-term neurodevelopmental outcomes of preterm infants.[7,8] In general, early detection of PGF is important to optimize nutritional support in neonates not growing well, with the aim of mitigating later adverse outcomes, such as neurodevelopmental impairment.[7,8]

Machine learning, an application of artificial intelligence using computer-based algorithms, has exhibited promising results in predicting clinical outcomes.[9] Supervised learning, in which computer algorithms are used to create models to assign input parameters from a dataset toward a preassigned outcome, is one commonly performed application of machine learning.[10] Owing to its outstanding reliability in processing complex relationships among various variables and producing stable predictions, machine learning has now been applied to several domains in the medical field, including prediction of disease progression or clinical outcomes.[10,11] For prediction of clinical outcomes, existing conventional risk models, such as logistic regression, have obvious limitations because they can only be applied to certain subsets of patients and require time-consuming, manual data entry.[9] Meanwhile, predicting the growth of preterm infants is not easy owing to the number of clinical variables involved, although recent machine learning techniques have been found to show promise in predictive models with good performance. To the best of our knowledge, a study on the prediction of PGF in neonates based on machine learning algorithms has not been conducted yet.

Thus, we aimed to develop a machine learning model with which to predict PGF among VLBW infants and to validate the performance of machine learning algorithms in comparison to the conventional multiple logistic regression (MLR) risk model.

MATERIALS AND METHODS

Study design and data collection

The KNN registry prospectively collects information about maternal antenatal and perinatal history, postnatal morbidities, growth outcomes, and other clinical outcomes of infants during hospitalization and long-term outcomes since their discharge until 3 years of age. The records of 10425 VLBW infants born between 2013 and 2017 and registered in the KNN were reviewed. Infants under a gestational age of 23 weeks or above 34 weeks of age and infants with severe congenital anomalies or who died before discharge were excluded. Data for 7954 infants were included for analysis. After excluding infants with missing values, 7954 VLBW infants were finally included in the study (Fig. 1). The KNN registry was approved by the Samsung Medical Center Institutional Review Board (2013-03-002), and the Institutional Review Boards of all 70 hospitals participating in the KNN. Written consent was obtained from the parents of infants during enrollment in the KNN. Data availability was subject to the Act on Bioethics and Safety [Law No. 1518, article 18 (Provision of Personal Information)]. Contact for sharing the data or accessing the data can be possible only through the data committee of the KNN (http://knn.or.kr) and after obtaining permission from the Centers for Disease Control and Prevention (CDC) of Korea.

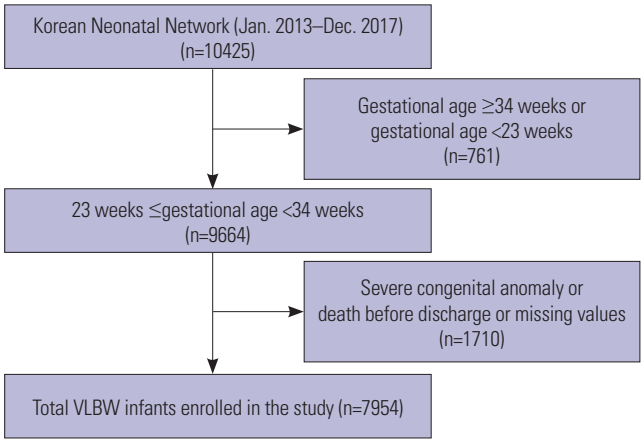

Fig. 1. Schematic flow chart of the enrolled very low birth weight (VLBW) infants.

The definitions were guided by the manual of operations of the KNN. The definition of maternal hypertension was defined as newly diagnosed hypertension in a pregnant woman after 20 weeks of gestation. Similarly, maternal diabetes mellitus included gestational diabetes mellitus and overt diabetes mellitus. Initial neonatal resuscitation was recorded when any of the following procedures was performed: oxygen supplementation, use of positive pressure ventilation, endotracheal intubation, cardiac massage, and administration of medications. Air leak syndrome was defined as a disease entity including pneumothorax, pneumomediastinum, pulmonary interstitial emphysema that required invasive procedures, such as the insertion of a chest tube or needle aspiration. Massive pulmonary hemorrhage was defined as pulmonary hemorrhage leading to cardiovascular collapse or acute respiratory failure. Pulmonary hypertension that required only pharmacological management, such as nitric oxide and sildenafil, was recorded. Bronchopulmonary dysplasia was defined as the need of oxygen use or the level of respiratory support at 36 weeks of postmenstrual age.[12] Necrotizing enterocolitis was defined according to modified Bell's criteria.[13] Sepsis was defined by blood cultures positive for bacteria or fungi and antibiotic therapy ≥5 days. Small for gestational age (SGA) was defined as birth weight lower than the 10th percentile for gestational age according to Fenton's growth chart.[14] Non-invasive ventilation was defined as any non-invasive positive pressure support, including continuous positive airway pressure and high flow nasal cannula. PGF was defined as a decrease in Z score of weight between birth

and discharge of more than -1.28 using Fenton's growth chart.[2,15]

Outcomes

The primary outcome was whether any machine learning algorithm showed better performance in predicting PGF at discharge than the classic MLR model. In addition, we checked specific time-points after birth that showed the highest effectiveness to predict PGF at discharge. We attempted to compare five time-points: at birth (Day 0), 7 days after birth (Day 7), 14 days after birth (Day 14), 28 days after birth (Day 28), and at discharge.

Owing to the myriad of variables in the KNN dataset, we selected those that had a statistical association with PGF through MLR analysis using a training dataset and redistributed the variables according to the corresponding time-points. Eventually, the variables for Day 0 were sex, gestational age, birth weight, SGA, maternal hypertension, and maternal premature rupture of membrane (PROM). Air leak syndrome, respiratory distress syndrome, intraventricular hemorrhage, and duration of invasive and non-invasive ventilation until 7 days after birth were added to the variables for Day 0 and set as the variables for Day 7. Medical and surgical treatments for patent ductus arteriosus were added to the variables for Day 7, and duration of invasive and non-invasive ventilation until 14 days after birth were changed instead of 7 days after birth and set as the variables for Day 14. Necrotizing enterocolitis ≥grade 2, spontaneous bowel perforation, and sepsis were added to the variables for Day 14, and duration of invasive and non-invasive ventilation until 28 days after birth were changed instead of 14 days after birth and set as the variables for Day 28. Finally, variables for at discharge were the same as those for Day 28, and the only difference was the actual duration of invasive and non-invasive ventilation.

Machine learning models

The following four machine learning models were used in the study: extreme gradient boosting (XGB), random forest (RF), support vector machine, and convolutional neural network.[16] We adopted 'Scikit learn 1.0.1, matplotlib 3.4.3 (AUC Graph), and the numpy 1.21' package in python 3.6. The dataset was composed of variables of short-term outcomes during the hospitalization of 7954 VLBW infants. We randomly divided the dataset into training and test sets at a ratio of 4:1. To avoid overfitting of the training set, the validation set was randomly selected in the training set at a ratio of 4:1. Internal validation was performed to secure as much data as possible to build an optimal model with high predictive power. The n-fold test was not used because the number of data was judged to be large enough. The hyperparameters were tuned using a grid search, and the final validated training models were applied to the test set. For RF parameters, the max depth was 3, 6, 9, and 12, and 100, 200, and 500 trees were explored. As parameters of XGB, booster used gbtree and gblinear; learning rates 0.1, 0.2, and 0.3 were searched; and max depths of 3, 6, 9, and 12 were used. For the convolutional neural network model, a network consisting of one to five nodes and one to five hidden layers was used, and the learning rate was searched using three parameters: 0.0001, 0.001, and 0.01.

To evaluate the diagnostic performances of each model, six metrics, including the area under the receiver operating characteristic curve (AUROC), accuracy, precision, sensitivity, specificity, and F1 score, were measured in the test set. As AUROC is a widely-used index to describe the ability of a machine learning model to predict outcomes,[17] we used it as the primary performance metric. The metrics ranged from 0 to 1, with values closer to 1 indicating a better model.[17] The error rate of each model was also analyzed.

Statistical analysis

Categorical variables are expressed as a n (%) and continuous variables as a mean±SD. To compare the baseline demographics between the training and test sets, the chi-squared test for categorical variables and independent two-sample t-test for continuous variables were used. Diagnostic performance, including accuracy, precision, sensitivity, specificity, and F1 score, was calculated using optimal cut-off points based on the predicted probability from the machine learning models for improving performance as much as possible. Error rate was also calculated in the same way. The optimal cut-off point was set to the point where Youden's index (defined as sensitivity+specificity-1) was maximized. All results were calculated based on the prediction probability of optimal cut-off points. To compare diagnostic performance between the models, we used the bootstrap method, which means that 1000 datasets allowed for duplication were randomly extracted and analyzed. Through this, we could obtain standard errors of the differences between the models considering the dependency between each data point. *P*-values based on z-statistic, which was calculated using the standard error obtained through bootstrap, under 0.05 were considered statistically significant. The analysis was conducted using SPSS version 23.0 (IBM Corp., Armonk, NY, USA) and R package version 4.0.3 (http://www.R-project.org).

RESULTS

The baseline demographics of the training and test sets are shown in Table 1. Among 7954 VLBW infants, the mean±SD gestational age was 28.98 weeks (2.4); mean birth weight was 1121.6 g (254.1). The incidences of PGF at discharge and SGA at birth were 44.2% (3515) and 22.1% (1266), respectively. The number of mothers with hypertension and PROM was 1760 (22.1%) and 2875 (36.1%), respectively. Most of the variables did not show statistical differences between the two sets.

The predictive performances of the four machine learning models and the MLR model for PGF at discharge are shown

https://doi.org/10.3349/ymj.2022.63.7.640

Table 1. Baseline Demographics of the Training Set and Test Set (n=7954)

	Training (n=6363)	Test (n=1591)	*p* value
Variables for Day 0			
Sex	3169 (49.8)	825 (51.9)	0.143
Maternal hypertension	1433 (22.5)	327 (20.6)	0.085
Maternal PROM	2297 (36.1)	578 (36.3)	0.857
Gestational age (weeks)	28.99±2.4	28.99±2.4	0.515
Birth weight (g)	1123.20±252.8	1115.23±259.5	0.263
SGA	1016 (16.0)	250 (15.7)	0.804
Variables added for Day 7			
Air leak syndrome	187 (2.9)	43 (2.7)	0.615
Respiratory distress syndrome	5049 (79.4)	1263 (79.4)	0.976
Intraventricular hemorrhage	852 (13.4)	211 (13.3)	0.885
Duration of invasive ventilation until 7 days of age (days)	3.51±3.0	3.57±3.0	0.490
Duration of non-invasive ventilation until 7 days of age (days)	4.85±2.6	4.88±2.6	0.646
Variables added for Day 14			
PDA medical treatment	2139 (33.6)	560 (35.2)	0.357
PDA surgical treatment	659 (10.4)	162 (10.2)	0.700
Duration of invasive ventilation until 14 days of age (days)	5.54±5.7	5.67±5.7	0.396
Duration of non-invasive ventilation until 14 days of age (days)	8.56±5.3	8.57±5.3	0.931
Variables added for Day 28			
Idiopathic spontaneous bowel perforation	98 (1.5)	27 (1.7)	0.653
Sepsis	1202 (18.9)	340 (21.4)	0.030
Necrotizing enterocolitis (≥stage 2)	297 (4.7)	81 (5.1)	0.479
Duration of invasive ventilation until 28 days of age (days)	8.38±10.5	8.74±10.7	0.235
Duration of non-invasive ventilation until 28 days of age (days)	14.12±10.7	14.25±10.8	0.654
Variables added for at discharge			
Total duration of invasive ventilation (days)	11.88±19.6	13.04±21.1	0.048
Total duration of non-invasive ventilation (days)	19.47±18.5	19.38±18.7	0.862

PROM, premature rupture of membrane; SGA, small for gestational age; PDA, patent ductus arteriosus.
Data are presented as mean±standard deviation or n (%).

in Table 2. All results were calculated based on the prediction probability of optimal cut-off points, not a default value of 0.5 (Table 3). Using variables for Day 0, almost all metrics of the XGB algorithm, except sensitivity, seemed to show better performance than those obtained via the other models; however, there were no statistical differences, compared with MLR. Using variables for Day 7, XGB still showed better performance in the metrics of AUROC [0.74 (95% CI 0.71–0.76)], accuracy [0.68 (95% CI 0.66–0.70)], and F1 score [0.67 (95% CI 0.64–0.70)], compared with those obtained via the other models. Compared with the MLR model, AUROC [0.74 (95% CI 0.71–0.76) vs. 0.72 (95% CI 0.70–0.75), p=0.03], sensitivity [0.73 (95% CI 0.70–0.76) vs. 0.68 (95% CI 0.65–0.71), p<0.01], and F1 score [0.67 (95% CI 0.64–0.70) vs. 0.65 (95% CI 0.62–0.68), p=0.03] of XGB were significantly higher. With the variables of Days 14 and 28, and at discharge, there were no statistically significant higher values in AUROC among the four machine learning models, compared with MLR.

The AUROCs of the XGB model among different time-points during hospitalization were compared to determine which time-point was most suitable for prediction of PGF at discharge (Table 4). AUROC seemed to increase gradually as time progressed towards the end of hospitalization [0.72 (95% CI 0.69–0.74) of Day 0, 0.74 (95% CI 0.71–0.76) of Day 7, 0.74 (95% CI 0.72–0.76) of Day 14, 0.74 (95% CI 0.72–0.77) of Day 28, and 0.75 (95% CI 0.72–0.77) of at discharge]; however, there was no statistically significant difference among the time-points, except for the comparison with Day 0 with other time-points (Day 7, Day 14, Day 28, and at discharge).

DISCUSSION

Presently, PGF is one of the most important issues in preterm infants, especially for smaller or more immature infants.[1,3,18] Despite recent improvements in nutritional support and treatment of preterm morbidities, PGF still accounts for a high percentage in preterm infants, and its prevalence shows great variance within neonatal intensive care units.[3,18] While early individualized and intensive nutritional care is essential for preventing PGF,[3,4] it is important to predict infants at high risk of developing PGF early after birth for better long-term growth and

Table 2. Predictive Performances of Four Machine Learning Models and MLR for PGF at Discharge

Models	XGB	RF	SVM	CNN	MLR
Day 0					
AUROC	0.72 (0.69–0.74)	0.67 (0.65–0.70)*	0.66 (0.64–0.69)*	0.66 (0.63–0.68)*	0.71 (0.69–0.74)
Accuracy	0.66 (0.64–0.69)	0.63 (0.61–0.66)	0.62 (0.59–0.64)*	0.61 (0.59–0.64)*	0.66 (0.63–0.68)
Error rate	0.34 (0.31–0.36)	0.37 (0.34–0.39)	0.38 (0.36–0.41)*	0.39 (0.36–0.41)*	0.34 (0.32–0.37)
Precision	0.60 (0.57–0.64)	0.60 (0.56–0.63)	0.56 (0.52–0.59)*	0.55 (0.52–0.59)*	0.60 (0.56–0.63)
Sensitivity	0.73 (0.70–0.76)	0.56 (0.52–0.59)*	0.71 (0.67–0.74)	0.71 (0.68–0.75)	0.71 (0.68–0.75)
Specificity	0.61 (0.58–0.64)	0.70 (0.67–0.72)*	0.54 (0.51–0.57)*	0.53 (0.50–0.56)*	0.61 (0.58–0.64)
F1 score	0.66 (0.63–0.69)	0.58 (0.55–0.61)*	0.62 (0.59–0.65)*	0.62 (0.60–0.65)*	0.65 (0.62–0.68)
Day 7					
AUROC	0.74 (0.71–0.76)*	0.72 (0.70–0.75)	0.67 (0.64–0.69)*	0.70 (0.68–0.72)*	0.72 (0.70–0.75)
Accuracy	0.68 (0.66–0.70)	0.65 (0.63–0.68)	0.61 (0.59–0.64)*	0.66 (0.64–0.68)	0.67 (0.65–0.69)
Error rate	0.32 (0.30–0.34)	0.35(0.32–0.37)	0.39 (0.36–0.41)*	0.34 (0.32–0.36)	0.33 (0.31–0.35)
Precision	0.62 (0.59–0.65)	0.59 (0.55–0.62)*	0.55 (0.52–0.59)*	0.60 (0.56–0.63)*	0.62 (0.59–0.65)
Sensitivity	0.73 (0.70–0.76)*	0.78 (0.75–0.81)*	0.72 (0.69–0.75)*	0.74 (0.71–0.77)*	0.68 (0.65–0.71)
Specificity	0.63 (0.60–0.67)*	0.55 (0.52–0.59)*	0.53 (0.50–0.56)*	0.59 (0.56–0.62)*	0.66 (0.63–0.69)
F1 score	0.67 (0.64–0.70)*	0.67 (0.64–0.70)	0.63 (0.60–0.65)	0.66 (0.63–0.69)	0.65 (0.62–0.68)
Day 14					
AUROC	0.74 (0.72–0.76)	0.73 (0.71–0.76)	0.67 (0.65–0.70)*	0.71 (0.69–0.74)*	0.73 (0.70–0.75)
Accuracy	0.68 (0.66–0.70)	0.68 (0.66–0.70)	0.62 (0.59–0.64)*	0.66 (0.64–0.69)	0.68 (0.66–0.71)
Error rate	0.32 (0.30–0.34)	0.32 (0.30–0.34)	0.38 (0.36–0.41)*	0.34 (0.31–0.36)	0.32 (0.29–0.34)
Precision	0.62 (0.59–0.66)*	0.64 (0.60–0.67)*	0.56 (0.52–0.59)*	0.60 (0.56–0.63)*	0.67 (0.63–0.70)
Sensitivity	0.71 (0.68–0.74)*	0.68 (0.64–0.71)*	0.72 (0.69–0.76)*	0.77 (0.74–0.80)*	0.59 (0.55–0.62)
Specificity	0.65 (0.62–0.68)*	0.69 (0.66–0.72)*	0.53 (0.50–0.56)*	0.58 (0.54–0.61)*	0.76 (0.73–0.79)
F1 score	0.66 (0.64–0.69)*	0.66 (0.63–0.68)*	0.63 (0.60–0.66)	0.67 (0.64–0.70)*	0.62 (0.59–0.65)
Day 28					
AUROC	0.74 (0.72–0.77)	0.75 (0.72–0.77)*	0.69 (0.66–0.71)*	0.71 (0.69–0.74)*	0.73 (0.71–0.76)
Accuracy	0.70 (0.68–0.72)	0.70 (0.67–0.72)	0.62 (0.60–0.65)*	0.68 (0.65–0.70)	0.69 (0.67–0.71)
Error rate	0.30 (0.28–0.32)	0.30 (0.28–0.33)	0.38 (0.35–0.40)*	0.32 (0.30–0.35)	0.31 (0.29–0.33)
Precision	0.65 (0.62–0.68)	0.65 (0.62–0.69)	0.56 (0.53–0.59)*	0.63 (0.59–0.66)*	0.66 (0.63–0.70)
Sensitivity	0.71 (0.67–0.74)*	0.69 (0.66–0.72)*	0.75 (0.72–0.78)*	0.68 (0.65–0.72)*	0.62 (0.59–0.66)
Specificity	0.69 (0.66–0.72)*	0.70 (0.67–0.73)*	0.52 (0.49–0.55)*	0.67 (0.64–0.70)*	0.74 (0.71–0.77)
F1 score	0.68 (0.65–0.70)*	0.67 (0.64–0.70)*	0.64 (0.61–0.67)	0.65 (0.63–0.68)	0.64 (0.61–0.67)
At discharge					
AUROC	0.75 (0.72–0.77)	0.75 (0.73–0.77)*	0.69 (0.67–0.72)*	0.72 (0.70–0.75)*	0.74 (0.71–0.76)
Accuracy	0.70 (0.67–0.72)	0.70 (0.67–0.72)	0.65 (0.62–0.67)*	0.67 (0.65–0.69)*	0.69 (0.67–0.71)
Error rate	0.30 (0.28–0.32)	0.30 (0.28–0.33)	0.38 (0.35–0.40)*	0.32 (0.30–0.35)*	0.31 (0.29–0.33)
Precision	0.65 (0.62–0.68)	0.65 (0.61–0.68)	0.61 (0.57–0.65)*	0.60 (0.57–0.64)*	0.65 (0.62–0.69)
Sensitivity	0.69 (0.66–0.73)*	0.71 (0.68–0.74)*	0.57 (0.53–0.61)*	0.76 (0.73–0.79)*	0.67 (0.63–0.70)
Specificity	0.70 (0.67–0.72)	0.69 (0.65–0.72)*	0.71 (0.68–0.74)	0.59 (0.56–0.63)*	0.71 (0.68–0.74)
F1 score	0.67 (0.64–0.70)	0.68 (0.65–0.70)	0.59 (0.56–0.62)*	0.67 (0.65–0.70)	0.66 (0.63–0.69)

PGF, postnatal growth failure; XGB, extreme gradient boosting; RF, random forest; SVM, support vector machine; CNN, convolutional neural network; MLR, multiple logistic regression; AUROC, area under the receiver operating characteristic curve.
*p<0.05 (compared with MLR).

neurodevelopmental outcomes. However, there is still no definite method or model for PGF prediction in high-risk preterm infants. Our study is the first to predict PGF of preterm infants through machine learning using sufficient data in premature newborns for whom it is not easy to collect large amounts of data.

In our study, based on a national-level database, we confirmed that some machine learning techniques, such as XGB and RF, show non-inferiority in terms of performance for PGF prediction in preterm infants, compared with MLR, a conventional statistical model. Furthermore, with the variables for day 7 after birth, we found that an XGB model could predict PGF during hospitalization better than the conventional MLR model with statistical significance in terms of the major performance

https://doi.org/10.3349/ymj.2022.63.7.640

Table 3. Optimal Cut-Off Points of Four Machine Learning Models and MLR by Youden's Index

	XGB	RF	SVM	CNN	MLR
Day 0	>0.3986730	>0.4748020	>0.3694691	>0.3923501	>0.4053404
Day 7	>0.4296191	>0.3707643	>0.3646140	>0.3796651	>0.4227439
Day 14	>0.4245050	>0.4427885	>0.3605821	>0.3587566	>0.4829504
Day 28	>0.4307109	>0.4462974	>0.3507205	>0.4578726	>0.4578184
At discharge	>0.4497405	>0.4416529	>0.4327583	>0.4162702	>0.4523962

XGB, extreme gradient boosting; RF, random forest; SVM, support vector machine; CNN, convolutional neural network; MLR, multiple logistic regression.

Table 4. Comparison of AUROCs of the XGB Model among Different Time-Points for Predicting PGF at Discharge Shown as *P*-Values

	AUROC	*p* value				
		vs. Day 0	vs. Day 7	vs. Day 14	vs. Day 28	vs. at discharge
Day 0	0.72 (0.69–0.74)	Reference	0.0045	0.0031	0.0028	0.0004
Day 7	0.74 (0.71–0.76)	0.0045	Reference	0.6918	0.3205	0.0793
Day 14	0.74 (0.72–0.76)	0.0031	0.6918	Reference	0.4514	0.1292
Day 28	0.74 (0.72–0.77)	0.0028	0.3205	0.4514	Reference	0.3790
At discharge	0.75 (0.72–0.77)	0.0004	0.0793	0.1292	0.3790	Reference

AUROC, area under the receiver operating characteristic curve; XGB, extreme gradient boosting; PGF, postnatal growth failure.

metric AUROC. In addition to AUROC, the performance metrics sensitivity and F1 score of XGB also showed significantly higher values than those obtained via the MLR model, which support the finding that the machine learning model had a higher predictive power for PGF of VLBW infants, compared with the conventional method.

To set the predictive models with high accuracy for predicting the PGF of preterm infants, it was essential to select variables appropriately associated with PGF. For this, we initially used MLR analysis with approximately 30 clinical variables in the KNN dataset and finally selected the variables that had statistically significance with PGF in the analysis. Most of them are known to be directly or indirectly associated with PGF. Birth weight and gestational age, which are direct indicators of the degree of prematurity, are correlated negatively with PGF in VLBW infants,[2,5] and sex differences have been found to affect fetal growth patterns.[6] SGA infants are more likely to develop growth failure than appropriate for gestational age infants,[2,19] and the adverse effects on growth last for years.[20] Maternal hypertension is a well-known risk factor for fetal growth restriction associated with placental insufficiencies,[21] and maternal PROM is known to increase the risk of neonatal sepsis.[22] Preterm morbidities, such as respiratory distress syndrome, air leak syndrome,[2,5] intraventricular hemorrhage, necrotizing enterocolitis, sepsis,[3,23,24] PDA,[25] and a longer duration of respiratory support[2,5] have significant adverse effects on postnatal growth, mainly through various mechanisms related to nutrition. Eventually, our analysis was conducted with appropriate variables that could affect PGF, except for the lack of information about nutritional support due to the limits of the database.

Recently in the clinical field of neonatology, machine learning models have been used in efforts to diagnose several diseases and predict clinical outcomes or prognosis. Neonatal seizures can be accurately detected by incorporating a machine learning algorithm into conventional electroencephalography;[26] furthermore, increasing the use of fundus photography has enabled the diagnosis of retinopathy of prematurity through computer-based image analysis. This method has the advantage of facilitating a reduction in fatigue susceptibility and other biases, compared to what human doctors can handle directly.[27] Among febrile infants, high-risk babies with serious bacterial infections can be predicted using supervised learning models.[28] We were also able to find preterm infants at risk of intracerebral bleeding using an RF model[29] and predict mortality in neonatal hypoxic-ischemic encephalopathy[30,31] or the perinatal period in developing countries.[32] Monitoring of vital signs in the neonatal intensive care unit (NICU) and early detection of adverse clinical outcomes are other examples of using machine learning techniques.[33] Machine learning is good for predicting clinical outcomes with limited data in newborns and infants for whom data are not easily obtained in large amounts in contrast to adults. In addition, ensembles of decision tree models, such as XGB and RF, may provide estimates of feature importance automatically from trained learning models.[34] From this, it is possible to infer why XGB could show better predictive power, compared with the conventional MLR model, in our study.

This study had some limitations. First, as the database created by collecting data from various NICUs nationwide, protocols for overall treatment and nutritional support of preterm infants could not be unified and controlled for the analysis. The lack of more detailed and informative data on nutrition, such as the types and timing of enteral feeding, use of fortifiers, overall periods, and compositions of parenteral nutrition, in the KNN database was also a major limitation. In addition, since the significant differences in some metrics between the machine learning algorithms and the conventional MLR model were

not large, we could not conclude that the machine learning algorithms, such as XGB, were a better way to predict PGF in VLBW infants than the conventional MLR. However, despite the limitations listed above, it is of immense significance because we demonstrated the possibility of a predictive model for PGF of preterm infants using machine learning. This is one of the few neonatal studies that has collected a lot of data from premature infants to obtain some meaningful results using machine learning algorithms. The development of machine learning algorithms that can achieve results while reducing unnecessary time and effort for collecting, organizing, and analyzing data, which are limitations of the conventional MLR model, is meaningful enough for both researchers and clinicians.

Using a nationwide preterm database, we confirmed machine learning models could predict PGF during the hospitalization of VLBW infants. Their use can help neonatologists diagnose high-risk preterm infants earlier in order to apply intervention to improve growth at an earlier period during NICU admission.

ACKNOWLEDGEMENTS

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT, the Ministry of Trade, Industry and Energy, the Ministry of Health & Welfare, the Ministry of Food and Drug Safety) (Project Number: KMMDF_PR_20200901_0057). Also, the authors thank Dong Hyun Kim InVisionLab, Inc. CEO, Artificial Intelligence Research Institute for assisting with the machine learning analyses.

AUTHOR CONTRIBUTIONS

Conceptualization: Jung Ho Han, Ho Seon Eun, and Soon Min Lee. Data curation: So Jin Yoon. Formal analysis: Hye Sun Lee and Goeun Park. Funding acquisition: Soon Min Lee. Investigation: Joohee Lim and Jeong Eun Shin. Methodology: Jung Ho Han, So Jin Yoon, Hye Sun Lee, and Soon Min Lee. Project administration: Soon Min Lee. Resources: Jung Ho Han and Soon Min Lee. Software: Dong Hyun Kim and Joohee Lim. Supervision: Min Soo Park. Validation: Joohee Lim and Jeong Eun Shin. Visualization: Goeun Park and Hye Sun Lee. Writing—original draft: Jung Ho Han and Soon Min Lee. Writing—review & editing: Ho Seon Eun, Min Soo Park, and Soon Min Lee. Approval of final manuscript: all authors.

ORCID iDs

Jung Ho Han https://orcid.org/0000-0001-6661-8127
So Jin Yoon https://orcid.org/0000-0002-7028-7217
Hye Sun Lee https://orcid.org/0000-0001-6328-6948
Goeun Park https://orcid.org/0000-0002-6670-5500
Joohee Lim https://orcid.org/0000-0003-4376-6607
Jeong Eun Shin https://orcid.org/0000-0002-4376-8541
Ho Seon Eun https://orcid.org/0000-0001-7212-0341
Min Soo Park https://orcid.org/0000-0002-4395-9938
Soon Min Lee https://orcid.org/0000-0003-0174-1065

REFERENCES

1. Horbar JD, Ehrenkranz RA, Badger GJ, Edwards EM, Morrow KA, Soll RF, et al. Weight growth velocity and postnatal growth failure in infants 501 to 1500 grams: 2000-2013. Pediatrics 2015;136:e84-92.
2. Lee SM, Kim N, Namgung R, Park M, Park K, Jeon J. Prediction of postnatal growth failure among very low birth weight infants. Sci Rep 2018;8:3729.
3. Griffin IJ, Tancredi DJ, Bertino E, Lee HC, Profit J. Postnatal growth failure in very low birthweight infants born between 2005 and 2012. Arch Dis Child Fetal Neonatal Ed 2016;101:F50-5.
4. Ziegler EE, Thureen PJ, Carlson SJ. Aggressive nutrition of the very low birthweight infant. Clin Perinatol 2002;29:225-44.
5. Lima PA, Carvalho Md, Costa AC, Moreira ME. Variables associated with extra uterine growth restriction in very low birth weight infants. J Pediatr (Rio J) 2014;90:22-7.
6. Alur P. Sex differences in nutrition, growth, and metabolism in preterm infants. Front Pediatr 2019;7:22.
7. Ehrenkranz RA, Dusick AM, Vohr BR, Wright LL, Wrage LA, Poole WK. Growth in the neonatal intensive care unit influences neurodevelopmental and growth outcomes of extremely low birth weight infants. Pediatrics 2006;117:1253-61.
8. Franz AR, Pohlandt F, Bode H, Mihatsch WA, Sander S, Kron M, et al. Intrauterine, early neonatal, and postdischarge growth and neurodevelopmental outcome at 5.4 years in extremely preterm infants after intensive neonatal nutritional support. Pediatrics 2009;123:e101-9.
9. MacKay EJ, Stubna MD, Chivers C, Draugelis ME, Hanson WJ, Desai ND, et al. Application of machine learning approaches to administrative claims data to predict clinical outcomes in medical and surgical patient populations. PLoS One 2021;16:e0252585.
10. Shamout F, Zhu T, Clifton DA. Machine learning for clinical outcome prediction. IEEE Rev Biomed Eng 2021;14:116-26.
11. Goto T, Camargo CA Jr, Faridi MK, Freishtat RJ, Hasegawa K. Machine learning–based prediction of clinical outcomes for children during emergency department triage. JAMA Netw Open 2019;2:e186937.
12. Jobe AH, Bancalari E. Bronchopulmonary dysplasia. Am J Respir Crit Care Med 2001;163:1723-9.
13. Walsh MC, Kliegman RM. Necrotizing enterocolitis: treatment based on staging criteria. Pediatr Clin North Am 1986;33:179-201.
14. Fenton TR, Kim JH. A systematic review and meta-analysis to revise the Fenton growth chart for preterm infants. BMC Pediatr 2013;13:59.
15. Lin Z, Green RS, Chen S, Wu H, Liu T, Li J, et al. Quantification of EUGR as a measure of the quality of nutritional care of premature infants. PLoS One 2015;10:e0132584.
16. Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020;9:14.
17. Shin D, Lee KJ, Adeluwa T, Hur J. Machine learning-based predictive modeling of postpartum depression. J Clin Med 2020;9:2899.
18. Cooke RJ. Improving growth in preterm infants during initial hospital stay: principles into practice. Arch Dis Child Fetal Neonatal Ed 2016;101:F366-70.
19. Pilling EL, Elder CJ, Gibson AT. Growth patterns in the growth-retarded premature infant. Best Pract Res Clin Endocrinol Metab 2008;22:447-62.
20. Monset-Couchard M, de Bethmann O, Relier JP. Long term outcome of small versus appropriate size for gestational age co-twins/triplets. Arch Dis Child Fetal Neonatal Ed 2004;89:F310-4.
21. Mateus J, Newman RB, Zhang C, Pugh SJ, Grewal J, Kim S, et al. Fe-

https://doi.org/10.3349/ymj.2022.63.7.640

tal growth patterns in pregnancy-associated hypertensive disorders: NICHD Fetal Growth Studies. Am J Obstet Gynecol 2019;221:635.e1-16.
22. Shane AL, Sánchez PJ, Stoll BJ. Neonatal sepsis. Lancet 2017;390:1770-80.
23. Clark RH, Thomas P, Peabody J. Extrauterine growth restriction remains a serious problem in prematurely born neonates. Pediatrics 2003;111:986-90.
24. Han JH, Yoon SJ, Lim JH, Shin JE, Eun HS, Park MS, et al. The impact of neonatal morbidities on child growth and developmental outcomes in very low birth weight infants: a nationwide cohort study. Eur J Pediatr 2022;181:197-205.
25. Hansson L, Lind T, Wiklund U, Öhlund I, Rydberg A. Fluid restriction negatively affects energy intake and growth in very low birthweight infants with haemodynamically significant patent ductus arteriosus. Acta Paediatr 2019;108:1985-92.
26. Pavel AM, Rennie JM, de Vries LS, Blennow M, Foran A, Shah DK, et al. A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial. Lancet Child Adolesc Health 2020;4:740-9.
27. Gensure RH, Chiang MF, Campbell JP. Artificial intelligence for retinopathy of prematurity. Curr Opin Ophthalmol 2020;31:312-7.
28. Ramgopal S, Horvat CM, Yanamala N, Alpern ER. Machine learning to predict serious bacterial infections in young febrile infants. Pediatrics 2020;146:e20194096.
29. Turova V, Sidorenko I, Eckardt L, Rieger-Fackeldey E, Felderhoff-Müser U, Alves-Pinto A, et al. Machine learning models for identifying preterm infants at risk of cerebral hemorrhage. PLoS One 2020;15:e0227419.
30. Abbasi H, Unsworth CP. Applications of advanced signal processing and machine learning in the neonatal hypoxic-ischemic electroencephalogram. Neural Regen Res 2020;15:222-31.
31. Slattery SM, Knight DC, Weese-Mayer DE, Grobman WA, Downey DC, Murthy K. Machine learning mortality classification in clinical documentation with increased accuracy in visual-based analyses. Acta Paediatr 2020;109:1346-53.
32. Mboya IB, Mahande MJ, Mohammed M, Obure J, Mwambi HG. Prediction of perinatal death using machine learning models: a birth registry-based cohort study in northern Tanzania. BMJ Open 2020;10:e040132.
33. Van Laere D, Meeus M, Beirnaert C, Sonck V, Laukens K, Mahieu L, et al. Machine learning to support hemodynamic intervention in the neonatal intensive care unit. Clin Perinatol 2020;47:435-48.
34. Kwon YS, Baek MS. Development and validation of a quick sepsis-related organ failure assessment-based machine-learning model for mortality prediction in patients with suspected infection in the emergency department. J Clin Med 2020;9:875.