Perinatology
pISSN 2508-4887 • eISSN 2508-4895

Original article
Perinatology Vol. 31, No. 4, December, 2020
https://doi.org/10.14734/PN.2020.31.4.209

Predictive Model for Motor Developmental Delay in Preterm Infants by Using Recurrent Neural Network

Seung Soo Kim, MD, PhD[1],
Jun Hwan Song, MD, PhD[1,2],
Ho Kim, MD, PhD[1,2]

[1]Department of Pediatrics, [2]Regional Newborn Intensive Care Center, Soonchunhyang University Cheonan Hospital, Cheonan, Korea

Objective: The aim of this study is to develop the predictive model for motor developmental delay in Korean preterm infants beyond neonatal intensive care unit.

Methods: The authors retrospectively investigated the medical records of premature infants who had undergone developmental test and discharged from the single regional newborn intensive care center. We collected 30 independent variables and the motor scale of the Korean version of Bayley scale of infant and toddler development III (K-Bayley III). The predictive modeling was conducted by 3 steps: 1) data preprocessing, 2) training predictive models, and 3) evaluation of final performance of each model. We used sensitivity as a primary evaluation metrics, and F1 score and area under precision-recall curve (AUPRC) as a secondary metrics.

Results: Total 359 subjects were enrolled in the study. Ten percent of subjects were below 80 in the motor scale (coding as '1' in the dependent variable). Recurrent neural network model showed the best performance (sensitivity 1.00, F1 score 0.36, AUPRC 0.22). XGBoost model (sensitivity 0.71, F1 score 0.63, AUPRC 0.65) and ridge logistic regression model (sensitivity 0.71, F1 score 0.56, AUPRC 0.60) also showed good performance.

Conclusion: Machine learning approach showed good predictive value for motor delay in Korean preterm infants. The further research by using big data from multicenter is needed.

Key Words: Premature infant, Machine learning, Child development, Clinical decision rules

Received: 20 November 2020
Revised: 9 December 2020
Accepted: 14 December 2020

Correspondence to
Ho Kim, MD, PhD
Department of Pediatrics, Soonchunhyang University Cheonan Hospital, 31 Suncheonhyang 6-gil, Dongnam-gu, Cheonan 31151, Korea

Tel: +82-41-570-2160
Fax: +82-41-572-4996
E-mail: c78141@schmc.ac.kr

서론

최근 수 십 년간 미숙아의 생존율은 비약적으로 향상되어 왔다.[1] 이러한 미숙아의 생존율 증가에 따라, 여러 가지 후유증을 가진 생존아들의 치료와 관리가 중요한 문제로 대두되기 시작했다.[2] 이 중에서도 신경 발달의 문제는 주요한 부분을 차지하고 있으며, 많은 의료 비용의 부담을 야기하고 있다.[3,4] 빠른 발견과 치료가 좋은 예후를 가져올 것으로 기대되고 있으며, 이에 발 맞춰 미숙아들의 신경 발달 후유증을 예측하기 위한 여러 연구들이 있었다.[5-7] 그동안 신경 발달 예후와 관련성이 보고된 변수들은 재태 연령, 출생 체중, 성별, 폐 이형성증(bronchopulmonary dysplasia, BPD), 가족의 사회경제학적 상태 등이 있다.[8,9] 하지만 이전 연구들은 대부분 미숙아의 생존과 중증도를 예측하기 위해 개발된 검사 및 척도를 이용하고 있으며, 위험인자의 발굴에 초점이 맞춰져 있었다. 또한 개발된 예측 모형들도 대부분 종속 변수와 독립 변수의 선형관계를 가정한 선형 회귀(linear regression) 혹은 로지스틱 회귀분석(logistic regression)을 사용하였으며, 비선형 문제에서도 예측력이 좋은 인공신경망(artificial neural network) 등의 기계학습(machine learning) 방법을 사용한 논문은 많지 않다.[6,7]

특히 우리나라에서는 미숙아들의 신경 발달과 관련된 위험인자에 대한 연구들은 있었지만, 아직까지 신경 발달 지연을 예측하기 위한 모형에 대한 연구는 없었다.[4,8,9] 아시아로 범위 확대를 해도 미숙아의 발달 지연 예측 모형에 대한 연구들은 많지 않은 실정이다. 우리나라와 인구학적 특성 및 의료 시스템이 다른 나라에서 개발된 예측 모형들을 그대로 사용하

는 경우에는 여러 가지 편향(bias)에 노출될 수밖에 없으며 정확한 예측이 어려워진다. 그러므로 국내 미숙아들에게 적용할 수 있는 신경발달 예후 예측 모형의 개발은 시급한 문제이다. 이에 본 저자들은 최근 여러 분야에서 좋은 성능을 보여주고 있는 인공신경망과 XGBoost (extreme gradient boost), 능형 로지스틱 회귀(ridge logistic regression) 알고리즘을 이용해서, 국내 미숙아들의 운동 발달 지연을 예측하기 위한 모형을 개발하고자 본 연구를 수행했다.

대상 및 방법

1. 대상

2016년 9월 1일부터 2019년 11월 30일까지 본원 신생아 집중치료실에서 입원 치료 후 퇴원해서, 교정연령 9개월에서 12개월 사이에 한국형 베일리영유아발달검사 3판(Korean version of Bayley scale of infant and toddler development, K-Bayley-III)검사를[10] 받은 출생 시 재태연령 37주 미만인 영유아 359명을 대상으로 했다. 본 연구는 순천향대학교 천안병원 임상시험심사위원회의 승인을 받았다(SCHCA2020-08-040).

2. 인구통계학적 특성 및 혈액 표지자, K-Bayley-III 검사 결과

대상군의 의무기록을 바탕으로 산모의 연령, 영유아의 성별, 태아 수, 조기양막파수 여부, 임신고혈압, 임신당뇨, 산전 스테로이드 투여, 산전 황산마그네슘($MgSO_4$) 투여, 재태연령, 원내 출생 여부, 출생체중, 출산 방법(질식분만 혹은 제왕절개), 아프가 점수(1분, 5분), 출생 체온, 신생아중환자실(neonatal intensive care unit, NICU) 입원기간, 전체 혈구계산(complete blood count, 검사에서 백혈구(white blood cell, WBC) 수, 적혈구크기분포폭(red blood cell distiribution width, RDW), 평균혈소판용적(mean platelet volume, MPV), 원내 출생아의 경우 출생 1시간 이내, 원외 출생아의 경우 입원 1시간 이내의 혈액가스분석(blood gas analysis)의 pH와 염기과잉(base excess, BE), 입원 중 동맥관 개존 진단 및 치료(약물 혹은 수술), 뇌초음파로 관찰된 뇌실내출혈(intraventricular hemorrhage), 뇌자기공명 영상으로 확인된 뇌실 주위 백질연화증(periventricular leukomalacia, Modified Bell's Staging 2단계 이상의 괴사성장염(necrotizing enterocolitis 및 이로 인한 수술 여부를 후향적으로 조사했다. 그리고 종속 변수인 교정연령 9개월에서 12개월 사이의 K-Bayley-III의 운동척도(motor scale) 결과도 의무기록에서 후향적으로 조사했다.

3. 기술적 통계 분석 및 예측모형의 개발과 평가

대상군을 K-Bayley-III 운동척도가 80점 이상인 정상군과 80점 미만인 경계선 및 발달 지연군으로 분류했다. 모형 학습 전 두 그룹의 재태 기간 및 출생, NICU 입원 과거력, 입원 시 검사실 검사 결과, 입원기간 등 변수들의 대표값과 통계적 차이를 분석했다. 연속형 변수들은 대표값을 중위수(사분위수 범위, interquartile range)로 표시했다. 일표본 Kolmogorov-Smirnov 검정을 사용하여 정규분포 여부를 분석하고, 정규분포를 따르는 연속형 변수들의 분석에는 독립표본 t검정(independent t test)을 시행했으며, 정규분포를 따르지 않는 연속형 변수들의 분석에는 독립표본 Mann-Whittney의 U검정(independent Mann-Whittney U test)을 사용했다. 범주형 자료들은 % (명)으로 표시하였으며, 카이제곱 검정 혹은 Fisher의 정확검정(기대 빈도가 5 미만인 칸이 25% 이상 존재하는 경우)을 사용해 분석했다. 예측모형의 개발은 1) 분석 자료의 전처리(data preprocessing)와 2) 기계학습(machine learning), 3) 예측 모형의 성능 평가의 3단계로 이루어졌다.

1) 전처리 과정에서는 우선, 변수의 결측치(missing value)에 multiple imputation by chained equations (MICE)를 이용하여 대치값을 채워 넣었다. 다음으로 전체 자료를 지연군 해당 여부에 따라 층화하여, 학습에 사용하는 훈련자료(training data set), 최적의 모형을 선택하기 위해 학습된 모형의 성능을 검증하는 검증자료(validation data set) 및 최종적으로 모형의 성능을 평가하는 평가자료(test data set)로 각각 6:2:2의 비율로 무작위로 나누었다. 다시 훈련자료의 평균과 표준편차를 이용하여 각 그룹들의 값을 정규화(normalization)했다. 마지막으로 지연과 정상간 비율의 불균형 문제(imbalanced data)를 해결하기 위해서, 훈련자료만 synthetic minority oversampling technique (SMOTE)를 이용하여 정상과 지연의 비율을 1:1로 보정했다.

2) 기계학습 단계에서 종속 변수는 지연군을 '1'로 정상군을 '0'으로 코딩한 후, 과거력 및 검사실 검사 결과를 독립 변수로 하여 예측모형을 만들었다. 예측모형은 능형 로지스틱회귀(ridge logistic regression)와 XGBoost (eXtreme Gradient Boost), 순환신경망(recurrent neural network, RNN)을 이용하여 만들었다. 각 알고리즘의 초매개변수(hyperparameter) 및 구조는 아래와 같이 결정하였다.

A. 능형 로지스틱회귀는 벌점의 강도를 조절하는 초매개변수 C를 100으로 설정하여 학습시켰다.

B. XGBoost는 5겹 교차검증(5 folds cross validation)을 이

용하여 최적의 초매개변수를 찾았다. 최종적으로 colsample bytree 0.6, gamma 0.5, learning rate 0.01, max depth 4, min child weight 1, n_estimators 200, subsample 1.0으로 초매개변수를 지정하여 모형을 학습시켰다.

C. RNN은 '입력층-bidirectional GRU-dropout-GRU RNN-dropout-batch normalization-출력층'의 순서로 구성했으며, 단계(time step)의 개수와 신경 절(node)의 개수는 각각 4개로 설정했다. Optimizer는 Adam (adaptive moment estimation)을 사용하였으며, 초매개변수는 learning rate 0.001, beta_1 0.9, beta_2 0.999, epsilon 1e-7, dropout 0.35, batch size 72, max epoch 2000으로 정했다.

3) 모델의 성능평가를 위한 평가 척도(evaluation metrics)로는 민감도(sensitivity, recall)를 주 척도로 사용했고, F1 score와 area under precision-recall curve (AUPRC)를 부 척도로 사용했다. 그 밖에 area under receiver-operating curve (AUROC), 정확도(accuracy), 특이도(specificity), 정밀도(precision, positive predictive value) 등도 구했다.

4. 통계 프로그램 및 패키지(package)

기술통계 분석에는 SPSS 25.0 (IBM Corp., Armonk, NY, USA)을 이용했으며, 분석 자료의 전 처리 및 모형의 학습과 평가에는 프로그램 언어 python 3.6과 패키지 tensorflow 2.6 및 sklearn 0.23.2, impyute 0.0.8, imblearn 0.7.0, matplotlib 3.2.1, pandas 0.25.3, numpy 1.16.4, xgboost 1.1.1를 이용했다.

결과

1. 기술적 통계(Table 1)

최종적으로 359명의 대상군이 분석과 기계학습에 포함되었다. 정상군과 지연군 사이에서 원내 출생 여부, 아프가점수(1분, 5분), MPV, 동맥관 개존의 치료(약물, 수술), 뇌출혈에서 통계적인 차이가 존재했다.

2. 예측모형의 최종 성능(Table 2, Fig. 1)

예측 모형의 최종 성능평가에서 RNN 모형이 민감도 1.00로 가장 좋은 성능을 보였으며, F1 score 0.36, AUPRC 0.22, AUROC 0.82, 정확도 0.65, 정밀도 0.22를 보였다. XGBoost로 만든 모형은 민감도는 0.71로 RNN에 비해 낮았지만, F1 score 0.63, AUPRC 0.65, AUROC 0.83, 정확도 0.92, 정밀도 0.56으로 민감도를 제외한 나머지 척도에서는 가장 우수한 성능을 보여주었다. 능형 로지스틱 회귀 모형은 민감도 0.71을 보였으며, F1 score 0.56, AUPRC 0.60, AUROC 0.81, 정확도 0.89, 정밀도 0.46으로 비교적 우수한 성능을 보였다.

3. 예측모형 구성에서 변수의 중요도(Fig. 2)

XGBoost 모형 구성에서 재태연령, 산모의 연령, 출생체중, 5분 아프가점수, MPV, BE, WBC, ICH, RDW, 태아의 수 등의 순서로 중요도가 높은 것으로 나타났다.

고찰

본 연구에서는 최근 기계학습 분야에서 이루어진 중요한 발전들을 적극 반영하여, 모형의 예측 성능을 높이기 위한 여러 가지 시도들을 했다. 제일 먼저 전체 연구 대상 중에 약 10%에 해당하는 지연군을 찾는 예측 모형을 개발하기 위해서 불균형 자료(imbalanced data)에 대한 대책을 준비했다. 이러한 불균형 자료를 그대로 이용하는 경우, 예측 모형들이 무조건 정상군으로 예측을 하여도 90% 이상의 정확도를 보이는 현상이 나타날 수 있다.[6,11] 이렇게 만들어진 예측 모형은 학습 과정에서는 성능이 좋은 것처럼 나타날 수 있지만, 실제 진료 현장에서 적용하는 것은 불가능하다.

우선 결측치(missing value)가 존재하는 대상을 제거하지 않고 MICE를[12] 이용하여 결측치에 대치값을 채워 넣었다. 많은 경우 주어진 자료에서 일부 대상에서 변수값이 누락되어 있을 수 있다. 충분히 큰 자료에서는 이러한 결측치가 포함된 대상을 분석에서 제외하는 경우도 있다. 하지만 규모가 작은 자료에서는, 특히 불균형 자료에서는 이렇게 결측치를 기준으로 대상을 제외시키는 경우 오류 발생의 위험이 커진다.[11] 이런 경우 본 연구처럼 분석 자료의 다른 변수값을 사용하여 결측치를 추정하고 대치하는 방법을 사용할 수 있다. 이 중 본 연구에서 사용한 MICE는[12] 다음 단계들을 거쳐서 이루어진다: 1) 우선 결측치들을 각 변수 별 평균값으로 대치한 후, 2) 한 변수에 대해서만 나머지 변수들을 이용하여 회귀식으로 예측치를 만들어서 대치한다. 3) 1)과 2)의 과정을 전체 결측치들을 대치할 때까지 반복하여 수행하여 한 주기(cycle)를 완성한다. 4) 수 차례의 주기를 반복하면서, 대치된 값을 갱신해 나간다.

다음으로는 학습과정에서 기계학습 방법을 이용한 과추출(oversampling) 방법인 SMOTE를[13] 이용하여 학습 자료에서의 지연군과 정상군의 비율을 1:1로 맞춰주었다. 이를 통해 모형들이 학습과정에서 불균형 자료에 의한 편향에 노출되지 않도록 의도했다.[11] 하지만 학습된 모형의 성능을 평가하는 과정에

Table 1. Patient Characteristics Stratified by Motor Scale of Korean Version of Bayley Scale of Infant and Toddler Development III (K-Bayely III)

Time step	Variables	Missing (%)	Motor scale ≥ 80 (n=323)	Motor scale < 80 (n=36)	*P*-value
	Age at test (months)		9.2 (8.7, 9.8)	8.7 (8.0, 9.6)	0.01
0	Age of mother (years)		33 (30, 36)	32.5 (28, 35)	0.14
	Sex (male)		53.9 (174)	52.8 (19)	0.90
	Number of fetus		1 (1, 2)	1 (1, 2)	0.73
1	PROM	3.3	25.1 (79)	34.4 (11)	0.25
	PIH	3.3	21.0 (66)	28.1 (9)	0.35
	Gestational DM	3.3	17.1 (54)	18.8 (6)	0.82
	Prenatal steroid	3.6	84.4 (266)	80.6 (25)	0.61*
	Prenatal $MgSO_4$	4.2	55.6 (174)	54.8 (17)	0.94
2	Gestational age (weeks)		33 (30.9, 34.7)	32.1 (29.1, 35.6)	0.63
	Inborn delivery		97.2 (314)	88.9 (32)	0.03*
	Birth weight (g)		1,830 (1485, 2165)	1,715 (1,155, 2,470)	0.51†
	Mode of delivery (c/sec)		91.6 (296)	88.9 (32)	0.53*
	Apgar score (1 minute)		6 (4, 7)	4 (2.5, 6)	0.01
	Apgar score (5 minutes)		8 (7, 9)	6 (5, 8)	<0.01
	1st body temperature (℃)		36.8 (36.7, 37)	36.8 (36.7, 37)	0.69
	1st BGA (pH)		7.3 (7.3, 7.4)	7.3 (7.3, 7.4)	0.63
	1st BGA (base excess)		-4.1 (-5.9, -2.2)	-3.8 (-5.8, -2.5)	0.86
	1st CBC (WBC, cells/L)		10.1 (7.4, 12.7)	9.3 (7.1, 13.1)	0.76
	1st CBC (MPV)		9.6 (9.2, 10.0)	10.1 (9.6, 10.3)	<0.01
	1st CBC (RDW)		16.1 (15.5, 17.2)	16.1 (15.5, 16.9)	0.71
3	RDS		45.5 (147)	55.6 (20)	0.25
	PDA		18.6 (60)	36.1 (13)	0.13
	Medication for PDA		18.3 (50)	33.3 (8)	0.03
	Ligation to PDA		3.1 (10)	13.9 (5)	0.01*
	BPD	0.3	6.2 (20)	25.7 (9)	<0.01*
	ICH	3.9	22.8 (71)	39.4 (13)	0.03*
	PVL	40.4	5.9 (11)	7.7 (2)	0.66*
	NEC	0.3	1.5 (5)	5.7 (2)	0.14*
	Surgery for NEC	0.3	0.0 (0)	2.9 (1)	0.10*
	Duration of admission		22 (13.5, 36)	32 (11.8, 57.8)	0.06

Continuous variables are presented as median (interquartile range) and statistical testing performed using Mann-Whitney U test (or independent t-test if variable has a normal distribution). Categorical variables are presented as % (n) and statistical testing performed using the chi-square test (or Fisher's exact test if the expected cell count was ≤5). The percentage of missing values is zero unless otherwise shown.
Abbreviations: PROM, premature rupture of membrane; PIH, pregnancy induced hypertension; DM, diabetes mellitus; c/sec, caesarian section; BGA, blood gas analysis; CBC, complete blood cell count; WBC, white blood cell; MPV, mean platelet volume; RDW, red cell distribution width; RDS, respiratory distress syndrome; PDA, patent ductus arteriosus; BPD, bronchopulmonary dysplasia; ICH, intracranial hemorrhage; PVL, periventricular leukomalacia; NEC, necrotizing colitis.
*Fisher's exact test; †Independent *t*-test.

Table 2. Model Performance to the Test Dataset Withheld during Training

Model	Sensitivity	F1-Score	AUPRC	AUROC	Accuracy	Specificity	PPV
Recurrent neural network	1.00	0.36	0.22	0.82	0.65	0.62	0.22
XGBoost	0.71	0.63	0.65	0.83	0.92	0.94	0.56
Ridge logistic regression	0.71	0.56	0.60	0.81	0.89	0.91	0.46

AUPRC, area under precision-recall curve; AUROC, area under receiver-operating curve; PPV, positive predictive value.

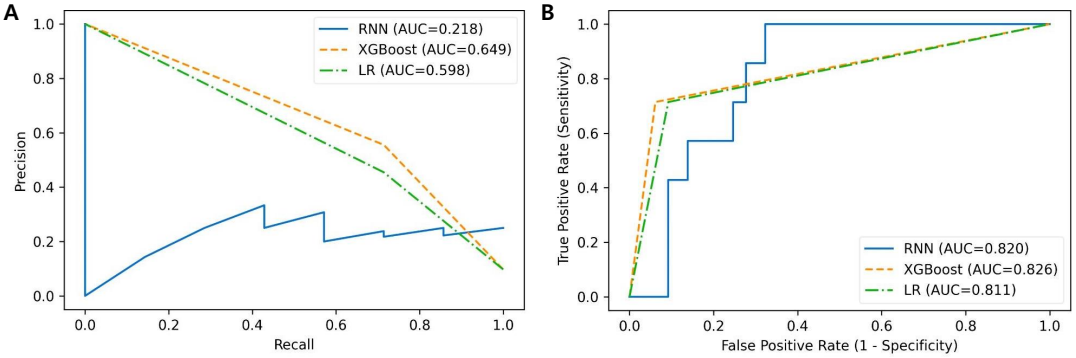

Fig. 1. These are precision-recall (PR) curve (A) and receiver operating characteristics (ROC) curve (B) of each predictive model. RNN, recurrent neural network; LR, ridge logistic regression; AUC, area under curve.

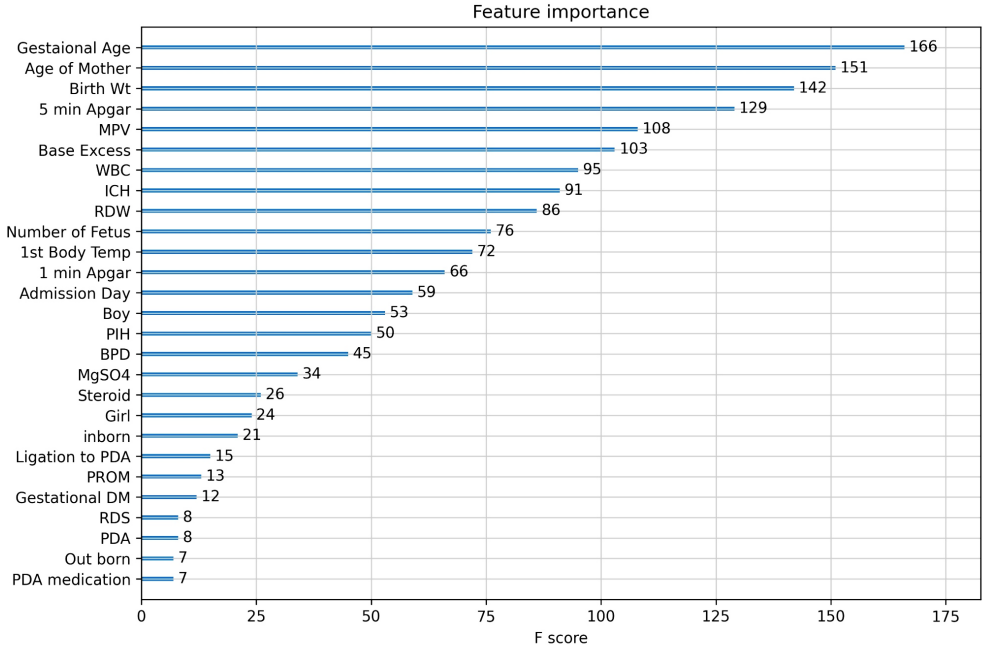

Fig. 2. These are relative importance of variables to the XGBoost model. Wt, weight; MPV, mean platelet volume; WBC, white blood cell; ICH, intracranial hemorrhage; RDW, red cell distribution width; Temp, temperature; PIH, pregnancy induced hypertension; BPD, bronchopulmonary dysplasia; inborn, inborn delivery; PDA, patent ductus arteriosus; PROM, premature rupture of membrane; DM, diabetes mellitus; RDS, respiratory distress syndrome; Out born, out born delivery.

서는 진료 현장에서 만나는 환자군과 유사한 자료를 사용해야 하기 때문에, 검증 자료와 평가 자료에는 이러한 SMOTE를 작용하지 않았다.

마지막으로 모형의 성능을 평가하기 위한 평가 척도(evaluation metrics)에 민감도와 AUPRC, F1 score를 사용했다. 민감도는 모형에서 예측한 지연군 중 실제 지연이 있었던 대상의 수(긍정 참, true positive, TP)를 실제 지연군의 수로 나눈 값이다. 본 연구에서 만든 모형의 목적은 발달 검사와 치료가 필요한 미숙아를 조기에 선별하는 것이기 때문에 민감도를 주 척도로 삼았다. 기계 학습에서는 민감도를 주로 재현율(recall)로 표기한다. 정밀도는 TP를 모형에서 예측한 지연군의 수로 나눈 값이다. 의학 분야에서 조금 더 익숙한 표현인 양성예측도(positive predictive value)와 동의어이다. AUPRC는 정밀도와 재현율의 상쇄 효과(trade off)를 나타낸 정밀도-재현율 곡선(precision-recall curve) 하단의 넓이이며, F1 score는 정밀도와 재현율을 곱한 것을 정밀도와 재현율로 나누고, 여기에 2를 곱한 값이다. AUPRC와 F1 score는 불균형 자료의 예측 모형의 성능 평가에 강점을 보인다.[6] 기계학습을 이용한 예측 모형의 평가에 흔히 사

용되고 있는 AUROC나 정확도 등을 사용하여 모형을 선택하는 경우에는 상술한 불균형 자료의 문제로 인해 우수한 성능의 모형을 찾을 수가 없다.[6,11]

본 연구에서는 예측 모형을 만들기 위해 기계학습의 방법을 사용했다. 고전 통계에서는 최소의 편향과 최소의 분산을 가진 최소분산불편추정량(minimum variance unbiased estimator) 모형을 구하는 것을 목표로 한다.[11,14] 하지만 이러한 모형은 현실에서는 분석에 사용한 데이터에서만 좋은 성능을 보이는 모형이 만들어지는 과적합(overfitting)에 빠지기 쉽고, 그 결과 새로운 자료에 대해서는 예측력이 떨어진다. 이와 달리 기계학습은 모형의 분산과 편향 사이의 상쇄(trade-off)를 고려하여, 검증자료에서 예측 오차를 최소화하는 모형을 추구한다.[11]

본 연구에서 종합적으로 가장 좋은 성능을 보여준 것은 XGBoost였다. 하지만 예측 모형의 목적이 발달검사를 대체하는 것이 아닌, 우선적으로 검사 및 치료가 필요한 대상을 선별하는 것이기 때문에 가장 높은 민감도를 보여준 RNN이 최종 예측 모형으로 선택되었다.

RNN은 인공 신경망의 한 종류이다. 인공 신경망은 뇌신경 세포(neuron)가 여러 다른 세포들로부터 받은 입력 신호의 총합(summation)이 역치(threshold)를 넘을 경우에만 다른 세포로 전달해주는 단순한 과정들이 모여서 복잡한 뇌의 활동이 가능해지는 것에서 아이디어를 얻어 1943년에 처음 제안되었다.[11,15] 이후 컴퓨터 성능의 한계로 실제 구현에 제약이 많아 관심을 받지 못하다가, 최근 컴퓨터 성능의 비약적 발전과 graphic processing unit에 의한 병렬 계산 도입에 힘입어 전성기를 맞이하게 되었다.[15] 특히 전통적인 선형 함수들로는 해결하기 어려운 비선형 문제나 추상적인 특성(abstractive representation)의 추출이 필요한 문제의 해결에 각광을 받고 있다. 이 중 RNN은 음성인식이나 기상 예측, 자연어 번역과 같은 시계열(time-series) 데이터의 처리에 좋은 성능을 보여주고 있다.[15] RNN은 기존 인공 신경망 모델과 달리 하나의 은닉 계층(hidden layer)이 다시 여러 단계(time step)를 가지고 있어서, 이전 단계에서 현재 단계에 필요한 정보들을 전달받을 수 있다.[15] 예를 들자면, 내일의 날씨를 예측하기 위해 오늘의 날씨에서 정보를 얻는 것이다. 이 중 본 연구에서 사용한 gated recurrent unit (GRU)는[16] long-short term memory (LSTM)과 같이 최근에 많이 쓰이고 있는 RNN의 한 종류로, 바로 이전 단계의 단기 기억 외에 여러 단계 이전부터 전달된 장기 기억도 이용할 수 있도록 설계되어 있다. 내일의 날씨를 예상하기 위해 오늘의 날씨뿐만 아니라 1주일 간의 날씨 정보를 이용하는 것이다. 하지만 GRU는 LSTM에 비해 조금 더 단순화된 구조로 계산 속도의 향상을 가져왔다. 그리고 Bidirectional GRU는[17] 이러한 GRU를 나란히 배치하여, 단계가 깊어질수록 변수들의 영향력이 약해져서 예측 모형의 성능이 떨어지는 기울기의 소실(vanishing gradient)을 해결하기 위해 고안된 알고리즘이다. 본 연구에서는 자료가 조사되는 시점을 기준으로 임의로 설정한 가상의 단계를 지정하여서, RNN 분석에 필요한 시간 축을 만들었다(Table 1).

XGBoost는[18] 최근 가장 많이 쓰이고 있는 부스팅(boosting) 알고리즘이다. 부스팅은 낮은 성능의 분류기(classifier)를 여러 개 조합해서 강력한 성능의 분류기를 만드는 방법으로, 분류기로 주로 결정 나무(decision tree)를 사용한다.[11] XGBoost는 랜덤 포레스트(random forest)에서[19] 처음 제안된 배깅(bootstrap aggregation, bagging) 기법을 차용하여, 매 학습 시마다 원래 자료에서 반복을 허용하는 무작위 추출로 분석 자료를 만들어서 사용한다. 현재 결정 나무에서 잘 분류하지 못했던 자료들을 다음 결정 나무에 사용할 분석 자료를 추출할 때 반영하고, 현재 결정 나무가 풀지 못한 어려운 문제들을 더 잘 풀 수 있도록 결정 나무를 만든다.[18] 그리고 각 결정 나무에 학습률(learning rate)을 적용하여, 과적합(overfitting) 문제를 해결하려는 시도를 하였다. 하지만 이러한 과정에서 우선 하나의 결정 나무가 만들어진 후에 다음 결정 나무가 만들어져야 하기 때문에, XGBoost 이전의 확률 경사 부스팅(gradient boosting) 방법들은 계산에 시간이 많이 걸린다는 치명적인 약점이 있었다. 하지만 2016년에 처음으로 제안된 XGBoost는 병렬 연산이 가능한 방법을 제안하여 계산에 필요한 시간을 단축시킬 수 있었고, 현재와 같이 널리 쓰이게 되었다.[18]

비교적 우수한 성능을 보여준 능형 로지스틱 회기는 벌점 회귀(penalized regression) 중 하나이다.[14] 여러 회귀모형에서 모형계수 벡터 β에 대한 최소제곱(least square) 해보다 이것을 약간 축소한 능형(ridge) 해를 쓰는 것이 예측 성능의 향상에 도움이 되는 것으로 알려져 있다.[11] 이를 일반선형화모형(generalized linear model)에 통합적으로 적용한 것을 정형화(regularization)라고 부르며, 일반선형회귀를 넘어서 로지스틱 회귀와 생존 분석 등에도 널리 적용되고 있다.[11] 이런 정형화는 설명 변수의 수가 대상군의 수에 비해 상대적으로 많은 경우에 특히 유효하다. 벌점화 회귀에서 λ는 초매개변수로 벌점(penalty)의 효과를 조절하며, λ가 커질수록 모형의 분산은 작아지고 편향은 커지는 편향분산 상쇄(bias-variance trade-off)를 보인다.[14]

본 연구에서는 그 동안 국내에서 연구된 적 없었던 미숙아들의 운동 발달 지연을 예측하기 위한 모형을 개발하였으며, 민감도와 AUROC, F1 score 등에서 이전 해외 연구들과 비교해서 손색이 없는 우수한 성능을 보여주었다.[7] 하지만 본 연구는 단일 기관의 후향적 자료를 이용하여 예측 모형을 만들었기 때문에, 다른 의료기관의 환자에 적용하기에는 여러 가지 제한점이 있

다. 추후 여러 기관에서 널리 사용할 수 있는 범용성이 높은 예측 모형을 만들기 위해서는 다기관 연구를 통해 여러 기관의 환자 특성을 반영한 빅데이터를 만들고, 이를 바탕으로 모형을 만들 필요가 있다. 또한 만들어진 모형은 전향적인 자료를 이용한 임상 검증을 통해 임상적 타당성(clinical validity)을 입증해야 할 것이다. 마지막으로 개별 기관에서 사용하기 전에 해당기관의 자료를 이용한 미세조정(fine tuning)을 거쳐서 사용 기관에서의 예측력을 높일 수 있도록 하는 것도 필요할 것이다.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

References

1) Lee JH, Youn Y, Chang YS. Short- and long-term outcomes of very low birth weight infants in Korea: Korean Neonatal Network update in 2019. Clin Exp Pediatr 2020;63:284-90.
2) Cho JY, Lee J, Youn YA, Kim SJ, Kim SY, Sung IK. Parental concerns about their premature infants' health after discharge from the neonatal intensive care unit: a questionnaire survey for anticipated guidance in a neonatal follow-up clinic. Korean J Pediatr 2012;55:272-9.
3) Mikkola K, Ritari N, Tommiska V, Salokorpi T, Lehtonen L, Tammela O, et al. Neurodevelopmental outcome at 5 years of age of a national cohort of extremely low birth weight infants who were born in 1996-1997. Pediatrics 2005;116:1391-400.
4) Jin JH, Yoon SW, Song J, Kim SW, Chung HJ. Long-term cognitive, executive, and behavioral outcomes of moderate and late preterm at school age. Clin Exp Pediatr 2020;63:219-25.
5) Spittle A, Orton J, Anderson PJ, Boyd R, Doyle LW. Early developmental intervention programmes provided post hospital discharge to prevent motor and cognitive impairment in preterm infants. Cochrane Database Syst Rev 2015(11):CD005495.
6) Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018;286:800-9.
7) Crilly CJ, Haneuse S, Litt JS. Predicting the outcomes of preterm neonates beyond the neonatal intensive care unit: What are we missing? Pediatr Res 2020 May 19;1-20 [Epub]. https://doi.org/10.1038/s41390-020-0968-5.
8) Choi SE, Lee KH. Analysis of risk factors for developmental delay in preterm infants using screening test. J Korean Child Neurol Soc 2018;26:146-51.
9) Kang JW, Lee KS. Prognostic factors of developmental delay in premature infants. J Korean Child Neurol Soc 2007;15:67-74.
10) Ahn SH, Yoo EY, Lee SH. A validation study of the gross motor scale of Korean version of bayley scales of infant and toddler development, third edition. Korean J Occup Ther 2018;26:81-97.
11) Kuhn M, Johnson K. Applied predictive modeling. 1st ed. New York (NY): Springer; 2013.
12) Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20:40-9.
13) Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. ICIC 2005;878-87.
14) de Vlaming R, Groenen PJ. The current and future use of ridge regression for prediction in quantitative genetics. Biomed Res Int 2015;2015:143712.
15) Gulli A, Kapoor A, Pal S. Deep Learning with TensorFlow 2 and Keras: Regression, ConvNets, GANs, RNNs, NLP, and more with TensorFlow 2 and the Keras API. 2nd Ed. Birmingham: Packt Publishing; 2019.
16) Chung J, Gülçehre Ç, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXivLabs 2014;abs/1412.3555.
17) Graves A, Fernández S, Schmidhuber J. Bidirectional LSTM networks for improved phoneme classification and recognition. Berlin: Springer; 2005.
18) Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016;785-94.
19) Breiman L. Random forests. Machine Learning 2001;45:5-32.