Review Article

Current progress and future prospects of machine learning in the diagnosis of neonatal encephalopathy: a narrative review

Yu-Fen Huang[1#], Zhong-Quan Jiang[2#], Lei Feng[3], Chao Song[4]

[1]Emergency Department, Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Centre for Child Health, Hangzhou, China; [2]School of Public Health, Lanzhou University, Lanzhou, China; [3]Department of Neonatology, Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Centre for Child Health, Hangzhou, China; [4]Department of Developmental and Behavioral Pediatrics, Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Centre for Child Health, Hangzhou, China

Contributions: (I) Conception and design: C Song; (II) Administrative support: C Song; (III) Provision of study materials or patients: YF Huang; (IV) Collection and assembly of data: YF Huang; (V) Data analysis and interpretation: ZQ Jiang, L Feng; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Chao Song, PhD. Department of Developmental and Behavioral Pediatrics, Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Centre for Child Health, 3333 Binsheng Road, Binjiang District, Hangzhou 310051, China. Email: songchao1987@zju.edu.cn.

Background and Objective: Neonatal encephalopathy (NE) can cause permanent neurological damage in newborns. NE greatly increases the burden of care placed on families. It also places a tremendous economic strain on the social health system. Currently, NE is mostly diagnosed by imaging and blood gas analysis. However, current diagnostic methods mostly lag behind the disease, leading to a lag in medical interventions for NE. In recent years, machine learning (ML) techniques have been applied to medicine, including in the early diagnosis and screening of diseases. This study aimed to provide an overview of existing research on the application of ML to NE and to offer insights for future investigations.

Methods: A full library search in fuzzy matching mode was performed to retrieve articles from the Web of Science database published between January 1, 2008, and August 31, 2024 using the following search strategy: (neonatal encephalopathy * machine learning) (where NE comprised all the relevant diseases, and ML comprised the main algorithms), and the key information was filtered.

Key Content and Findings: A total of 159 documents were retrieved, and 23 relevant documents were identified based on the topic, keywords and content. The relevant content showed that the included articles on NE and ML had issues in terms of study standardization, dichotomous study outcomes, and clinical usefulness.

Conclusions: To date, most studies on the application of ML to NE have not comprehensively considered the aspects of experimental design, data processing, model building, and evaluation. It is hoped that such models will provide effective decision-making tools for clinical practice in the future, and thus improve the healthy life span of newborns.

Keywords: Neonatal encephalopathy (NE); machine learning (ML); diagnosis; application

Submitted Oct 13, 2024. Accepted for publication Mar 25, 2025. Published online Apr 27, 2025.
doi: 10.21037/tp-24-425
View this article at: https://dx.doi.org/10.21037/tp-24-425

Introduction

Neonatal encephalopathy (NE) is a general term for a group of acute neurological disorders characterized by abnormal states of consciousness, muscle tone or reflexes, and convulsions that occur in newborns ≥35 weeks of gestational age (1). The main feature of NE is varying degrees of consciousness impairment (2). NE impedes the development of memory, language, cognition, behavior, and computation (3), and is a major cause of distant neurodevelopmental disorders in newborns. As a result, it severely reduces the healthy life expectancy and quality of life of newborns, especially in low- and middle-income countries (4). According to incomplete statistics, 8.5 cases per 1,000 live births are estimated to develop NE associated with intrapartum events, and the majority (96%) of these newborns are born in low- and middle-income countries (5). NE-related diseases diseases mainly include hypoxic-ischemic encephalopathy (HIE), asphyxia, intracranial hemorrhage (ICH), hyperbilirubinemia, cerebral palsy (CP), and intrauterine growth restriction (IUGR).

NE is merely a description of neurological function; thus, its etiology cannot be determined (6). In developing countries, HIE is the single most common cause of NE, which leads to an expanded diagnosis of NE. Furthermore, this expanded diagnosis of NE hinders etiology-based diagnosis and precise intervention. Therefore, the diagnosis of NE requires the careful use of imaging data, biochemical indicators, and the patient's clinical history by experienced pediatricians (7). This places a significant burden in underdeveloped areas (8).

Researchers have begun to use machine learning (ML) to optimize the diagnosis of NE-related disorders (9-12). ML can assist in NE diagnosis because it: (I) is able to process imaging data; (II) can solve covariance problems well; (III) has very high computational efficiency on high-performance computers; (IV) is able to output the importance ranking of features; and (V) has a superior ability to handle high-dimensional problems (13). The performance of ML models in establishing a diagnosis of NE is usually better than that of traditional logistic regression (LR) and Cox regression models (14-18). Efficient diagnostic tools are crucial for clinical intervention and the long-term prognosis prediction of NE. Thus, high-performance ML-based diagnostic tools for NE need to be established.

ML is a multidisciplinary field that includes probability theory, statistics, and complex algorithms (19). ML algorithms can extract information from existing data sets and make accurate predictions about unknown data. The problem of diagnosis in the field of NE-related diseases can be understood as an ML classification problem. This classification problem can be solved by training an automated classifier with features from a dataset classified as diseased and non-diseased to make diseased and non-diseased diagnoses using new data.

ML includes supervised, unsupervised, semi-supervised, and reinforcement learning (19). Supervised learning algorithms are often used in classification problems to assist in diagnosis. In supervised learning, common classification algorithms include LR, support vector machine (SVM), random forest (RF), eXtreme Gradient Boosting (XGBoost), and neural networks. The core problem of these algorithms in the model building process lies in the feature engineering of the underlying research data, including the normalization, standardization, category balancing, identification and complementation of missing values, and feature selection of data. In relation to the imaging data, the more critical issue is data feature extraction.

This study sought to review studies that applied ML to NE in recent years. It also critically evaluated the limitations and strengths of these studies. More specifically, the study (I) examined the application of ML methods; (II) considered the treatment of feature engineering; and (III) evaluated relevant predictive models.

Most existing reviews of NE studies have focused on the association between diseases and future research progress (20,21), and reviews of NE studies from the perspective of ML domain are limited. Thus, unlike other review studies, this study focused on data preprocessing, feature engineering, and the multidimensional evaluation of models from an ML perspective. In addition, it also examined whether the models or findings of previous studies could be integrated to build an early warning prediction model for NE that is applicable to most diseases. This model could lead to the establishment of an efficient early warning and prediction system for neonatal brain development that could be used in primary child health care in less developed areas.

This article begins by evaluating different existing ML studies on NE-related disorders. It then draws conclusions and outlines some future research directions. Finally, it draws conclusions and identifies some future research directions. We present this article in accordance with the Narrative Review reporting checklist (available at https://tp.amegroups.com/article/view/10.21037/tp-24-425/rc).

Table 1 The search strategy summary

Items	Specification
Date of search	August 31, 2024
Database searched	Web of Science
Search terms used	Neonatal Encephalopathy * Machine Learning
Timeframe	2008/01/01–2024/08/31
Inclusion criteria	(I) NE studies using machine learning methods with a population of newborns; and (II) English-language articles
Selection process	To ensure the accuracy of the investigation, C.S. and Z.Q.J. independently screened the retrieved papers under supervision. They assessed the titles and abstracts to determine the eligibility of articles in relation to predetermined inclusion and exclusion criteria. If disagreements arose between the two reviewers, internal discussions ensued to facilitate consensus. If the disagreement continued, the final decision was made by a third reviewer (C.S.)

NE, neonatal encephalopathy.

Methods

A full library search in fuzzy matching mode was performed to retrieve articles from the Web of Science database published between January 1, 2008, and August 31, 2024 using the following search strategy: (Neonatal Encephalopathy * Machine Learning) (where NE comprised all the relevant diseases, and ML comprised the main algorithms), and the key information was filtered. A total of 159 documents were initially retrieved, and 23 relevant documents were identified based on the topic, keywords, and content (see *Table 1*).

With recent advances in medical diagnostic technology, imaging data, such as cranial ultrasound, magnetic resonance imaging (MRI) and electroencephalogram (EEG) data, have been widely used in the diagnosis of diseases related to NE. In addition, the birth history and maternal factors of newborns cannot be ignored in the diagnosis of diseases related to NE. Using these data, researchers have used different ML methods to build diagnostic models for NE-related diseases. In the following sections of this article, current applications of ML in the field of NE are delineated based on different disease perspectives.

Asphyxia and neonatal HIE

Asphyxia and HIE are the most common causes of NE. The diagnosis of HIE depends on the obstetric history, Apgar score, and neurological symptoms. In addition, EEG, ultrasound, computed tomography, or MRI can also assist in the diagnosis of HIE to some extent. Usually, to make a diagnosis of HIE, a complete examination must be performed to determine the pathological type and clinical classification. Mooney *et al.* (22) sought to help clinicians to predict the occurrence of HIE in neonates early and accurately, and used an RF model to develop an early warning prediction model for HIE in infants with perinatal asphyxia (PA). The characteristics included the birth history of the newborn, blood gas analysis, and maternal data. The feature selection method used the iterative RF model, and the top five important predictors were found to be the infant's condition at birth (as expressed by the Apgar score), the need for resuscitation, and the first postnatal measures of potential of hydrogen (pH), lactate, and the base deficit. A similar approach was adopted by Pavel *et al.* (23), who developed a predictive model for seizures within 12 hours of birth in HIE infants using clinical and EEG parameters. In terms of the performance metrics, the quantitative amplitude-integrated EEG model (n=159) had a Matthews correlation coefficient (MCC) of 0.381, and an area under the curve (AUC) of 0.696, while the clinical and quantitative amplitude-integrated EEG model had an MCC of 0.384 and an AUC of 0.720. In addition, the investigators sought to exclude blood gas analysis data to improve the efficiency of the model in clinical practice. The model for the three different subtypes without blood gas analysis data had an AUC of 0.84–0.89, while the model with blood gas analysis data had an AUC of 0.80–0.98. Thus, the HIE prediction model without blood gas analysis data could also be used in clinical practice in terms of differentiation in HIE infants.

Usually, HIE can be diagnosed based on a history of birth asphyxia and clinical signs. Few investigators have combined metabolic and clinical data to predict HIE. Notably, O'Boyle *et al.* (14) combined clinical and metabolic data, and used LR and RF methods to predict HIE. They

found that the model that combined clinical and metabolic data had a discrimination AUC of 0.96 [95% confidence interval (CI): 0.92–0.95].

ML models for HIE diagnosis using electroencephalographic and MRI imaging data typically have higher performance compared to traditional diagnostic methods (24-28), and can more accurately assess the severity of HIE (24). However, the processing of diagnostic data often requires sophisticated feature extraction methods, such as the apparent diffusion coefficient Z-scores (Z_{ADC}) measurement (25) and principal component analysis (PCA) (26). Unlike Raurale *et al.* and Zheng *et al.* (24,26), Weiss *et al.* (25) focused on MRI and clinical data from children with established HIE, and employed ML methods to develop a hierarchical diagnostic model. They reported that the performance of their model was comparable to that of experts.

ICH

ICH most often occurs in preterm infants born at less than 32 weeks of gestation. Turova *et al.* (29) used the blood gas analysis data of 229 preterm infants and maternal data to build an RF model. Feature selection was performed using recursive feature elimination (RFE), and five blood gas analysis features were selected for subsequent modeling. During the modeling process, the researchers focused on the problem of data imbalance in cross-validation. The study also focused on the differences between preterm and very preterm infants, and constructed separate diagnostic ICH models for each group. The best model had an accuracy of 0.959 (95% CI: 0.885–0.991) for preterm infants and 0.972 (95% CI: 0.855–0.999) for very preterm infants, and both models had a sensitivity and specificity above 0.9. From a discriminatory perspective, these models were able to use blood gas analysis data for ICH prediction in preterm infants. Notably, this study not only explored the performance of the model from a discriminatory perspective, but also simultaneously explored the calibration of the model using calibration plots. This comprehensive analysis enhances its value for future clinical applications.

Jin *et al.* (15) developed a model to predict the long-term versus short-term prognosis of ICH patients using claims data. First, short-term prognosis was compared using five ML methods, and RF was ultimately selected for modeling. The best RF model had an AUC of 0.882. In terms of rare neonatal cerebral hemorrhage models, Guedalia *et al.* (30) built a subgaleal hemorrhage prediction model using maternal and fetal variables collected during the first stage of labor. Among the three integrated algorithms examined [i.e., balanced random forest (bRF), CatBoost, and AdaBoost], bRF showed the best predictive performance (AUC: 0.88, 95% CI: 0.856–0.904). Similarly, Kim *et al.* (11) used ultrasound to predict the occurrence of germinal matrix hemorrhage-intraventricular hemorrhage. They highlighted that their convolutional neural network (CNN) model exhibited exceptional performance, achieving an AUC of 0.92.

Hyperbilirubinemia, CP, and IUGR

In the diagnosis of hyperbilirubinemia, more variables have emerged that differ from clinical and ancillary diagnostic data, such as gut microbiota (31), genetic features (12), and gut metabolites (32). The presence of these variables has been linked to the etiology of hyperbilirubinemia. For example, Zhang *et al.* (31) used an RF model to construct a diagnostic model for neonatal jaundice using gut microbial data, and reported that the model had an AUC of 0.969 (95% CI: 0.904–1.000). Similarly, Zeng *et al.* (32) used microbial metabolic data to build an RF diagnostic model that had an AUC of 0.969, and Deng *et al.* (12) used genetic features and clinical risk factors to build a gradient boosting decision tree (GBDT) model that had an AUC of 0.795 (95% CI: 0.761–830). The above studies illustrate that models built using gut microbial and metabolic data are significantly better than those built using genotyping data. Notably, both the gut microbe models appeared to be overfitted.

Chou (33) further incorporated the effect of time on hyperbilirubinemia into a model. A long short-term memory (LSTM) network model was used to focus on the role of time series. Feature selection was performed using RF and XGBoost models to screen for the following seven variables: current bilirubin measurement, last rate of rise, proportion of time under phototherapy, time to next measurement, gestational age at birth, current age, and fractional weight change from birth. Subsequently, they compared LR, RF, multilayer perceptron (MLP), and LSTM with XGBoost, and found that MLP had the best discrimination ability (AUC: 0.94, 95% CI: 0.91–0.97).

Genotyping and general movement assessments are the key predictor variables for CP. Bahado-Singh *et al.* (34) used genotyping data to select the best model among four ML models [i.e., deep learning (DL), SVM, RF, and LR] for predicting CP. They found that the DL model had the best discrimination ability (AUC: 0.976, 95% CI: 0.676–1.00, sensitivity: 0.95, specificity: 0.944). The CP prediction

model was built using LR only and the general movement assessment. The final model had an AUC of 0.74.

Additionally, Pini *et al*. (35) used SVM to build a prediction model that had an accuracy of 0.93, a sensitivity of 0.93, and a specificity of 0.84. Signorini *et al.* (36) built an RF model to predict IUGR that had an AUC of 0.911 (95% CI: 0.860–0.961).

CardioTocographic recording features are often extracted using methods such as the RFE technique, the fetal heart rate-based encompassing time domain, frequency, and non-linear domain approaches. Sufriyana *et al.* (37) used preeclampsia and IUGR as outcomes of placental dysfunction-related disorders. The variables of the model included maternal characteristics, uterine artery (UtA) Doppler measures, and other information. The feature selection method used a correlation-based approach and a backward greedy stepwise search. The best RF model had an AUC of 0.976. Notably, this study used the precision-recall curve (PRC) to examine the discrimination ability of the model on an unbalanced data set (PRC: 0.958).

ML in NE: dilemmas and solutions of ML in NE applications

Currently, most of the aforementioned studies claimed that they used ML models that can be applied to clinical studies. However, in fact, most of the studies only applied existing ML methods to NE-related disease data. The main implementation processes of the existing literature are summarized in *Table 2*. Most of the studies had certain shortcomings related to the processing of data, the selection of methods, and the evaluation of models. Such shortcomings make it difficult to apply these models in clinical practice. In the third part of this article, the current problems and application solutions will be discussed in more detail.

Concepts and reflections

Standardize the modeling process

A complete report of the development and validation process of clinical diagnostic prediction models is essential for the external validation and clinical application of models. The 22 studies on ML prediction models for NE published in the last three years suffer from problems such as an insufficient amount of data, the improper handling of missing values, model construction strategy problems, and a lack of validation.

ML models do not have a stringent requirement in terms of the sample size. However, in small sample size training tasks, such models are highly susceptible to the overfitting of small samples or the underfitting of task objectives. Jeong *et al*. and Bahado-Singh *et al*. (10,34) noted that ML methods are applicable to large samples, but RF and DL were still used in very small samples. The final discrimination AUC for both RF and DL models was above 0.970. In addition, Jeong *et al.* (10) did not report the CI of the AUC, while that reported by Bahado-Singh *et al.* (34) was 0.676–1.000. The above RF and DL models can be considered to have an overfitting problem to some extent. A review by Lu *et al.* (39) had three suggestions for small sample size learning: (I) increase the training data; (II) reduce the space that the model needs to search; and (III) optimize the process of searching the model (39). Conversely, models with excellent performance on small sample tasks, such as SVM, could also be chosen (40).

The presence of missing values is inevitable in clinical practice. However, among the above-mentioned studies, only Chou *et al.* and Bahado-Singh *et al.* (33,38) reported on the treatment of missing values. Chou *et al.* (33) used median imputation, while Bahado-Singh *et al.* (38) used 0 for imputation. These value imputation methods can reduce the waste caused by removing samples to some extent. The method of imputing with 0 is the easiest to implement, but its drawback lies in the interference it may cause to the sample (e.g., in regression or classification, the estimates of the parameters may show greater deviation from the true values). Consequently, some researchers have proposed the great likelihood method or multiple imputation methods (41). The extreme likelihood imputation method is suitable for large samples. However, in most cases researchers prefer the multiple imputation method. The ideas underlying multiple imputation and Bayesian estimation are consistent; however, multiple imputation addresses the shortcomings of Bayesian estimation and retains the advantage of a single imputation. Conversely, multiple imputation requires more effort in data analysis. It is also reasonable to leave the missing values untouched and deal with the features directly. For example, Mooney *et al.* (22) observed a 20% missing problem in cord pH, but did not use imputation to deal with this feature; rather, they considered the need for efficient and readily available data for HIE prediction in clinical practice, and directly removed this feature from the model.

According to the "no free lunch" theorems in the field of ML (42), the application of algorithms should depend on

Table 2 The main implementation processes adopted by the existing studies

Study	Year	Disease selected	# of outcomes being divided	Data size and source	Sample division	Feature type	ML methods used	Feature selection	# of features being selected	CV	Balancing	Best algorithm	Classification	Calibration	Clinical usefulness	External validation
(22)	2021	HIE	4	68 mild HIE, 44; moderate HIE, 17; severe HIE 79 PA	7:3	Birth history of the newborn; maternal factors	RF	Iterative RF	Model 1: 12; Model 2: 10	5-fold	No	RF	Model 1: mod/severe HIE *vs.* PA/mild HIE AUC: 0.84; Model 2: mod/severe HIE *vs.* PA/mild HIE AUC: 0.94	No	No	No
(23)	2023	HIE	2	162 infants with HIE (53 had seizures)	No	Birth history of the newborn; maternal factors; EEH	RF and gradient boosting algorithms	Gradient boosting algorithms	13	10-fold	No	RF	AUC: 0.832	No	No	No
(27)	2023	HIE	2	186 HIE patients, 219 healthy controls	8:2	MRI	DLCRN	No	No	No	No	DLCRN	Internal validation: AUC: 0.813. External validation: AUC: 0.798	No	Yes	Yes
(28)	2024	HIE	3	414 neonates, of whom 198 died at 2 years	8:1:1	MRI	CNN	No	No	No	No	CNN	AUC: 0.77 (95% CI: 0.63–0.90)	No	No	Yes
(9)	2017	Asphyxia	2	31 very preterm infants	No	EEG	LR, LDA, KNN, SVM, Th	No	No	No	No	KNN, SVM, LR	Accuracy: 0.95	No	No	No
(24)	2021	HIE	4	54 term infants	No	EEG	CNN, GMM	No	No	5-fold	No	CNN	Accuracy: 0.89	No	No	Yes
(14)	2021	HIE	2	41 HIE, 40 PA, 40 healthy controls	No	Metabolites, birth history of the newborn	RF, LR	LASSO	Only metabolites: 5; only neonatal birth history: 3; integrated model: 4	10-fold	No	RF	AUC: 0.96 (95% CI: 0.92–0.95)	No	No	No
(25)	2019	HIE	-	300 term infants	No	MRI, clinical information	SVM, RF, LR	Z_{ADC} measurement	No	10-fold	No	No	No	No	No	No
(26)	2020	HIE	2	18 controls, 40 HIE/MRI– patients, 15 HIE/MRI+ patients	No	MRI	No	PCA	No	1-fold	No	No	No	No	No	No
(10)	2022	NE	3	24 term infants	No	MRI	RF	DWIC, FBA	No	2-fold	No	RF	Accuracy: 0.987	No	No	No
(15)	2022	ICH	3	5,926 neonates with ICH	7:3	Birth history of the newborn, maternal factors	LR, RF, SVM, GBDT	Backward selection in LR	4	No	No	RF	AUC: 0.882	No	No	No
(30)	2022	SGH	2	SGH following vacuum extraction n1=2,955; control n2=35,552	No	Birth history of the newborn. maternal factors; neonatology diagnosis	bRF, CatBoost, AdaBoost	No	No	3-fold	No	bRF	AUC: 0.88 (95% CI: 0.856–0.904)	No	No	No
(11)	2022	ICH	2	400 neonates with ICH	6:2:2	Ultrasound	CNN	No	No	No	No	CNN	AUC: 0.92	No	No	No
(29)	2020	ICH	2	118 patients with ICH, 111 patients without ICH	No	Birth history of the newborn, maternal perinatal factors, neonatology diagnosis	RF	RFE	5	10-fold	SMOTE	RF	Accuracy: 0.959 (95% CI: 0.885–0.991)	No	No	No
(33)	2020	Hyperbilirubinemia	2	38,748 newborn infants	9:1	Birth history of the newborn, maternal perinatal factors, neonatology diagnosis	LR, RF, MLP, LSTM, XGBoost	RF, XGBoost	8	10-fold	No	MLP	AUC: 0.94 (95% CI: 0.91–0.97)	No	No	No
(31)	2022	Hyperbilirubinemia	2	138 newborn infants: 69 neonates with jaundice, 69 healthy controls	No	Gut microbiota	RF	No	No	No	No	RF	AUC: 0.969 (95% CI: 0.904–1.000)	No	No	No
(12)	2021	Hyperbilirubinemia	2	984 newborn infants	7:3	Genetic features, clinical risk factors	GBDT, Cart, LR, RF	No	No	10-fold	No	GBDT	AUC: 0.795 (95% CI: 0.761–830)	No	No	No
(32)	2022	Hyperbilirubinemia	2	68 neonates with jaundice; 68 healthy controls	6:4	Gut metabolites	RF	No	No	No	No	RF	AUC: 0.969	No	No	No

Table 2 (*continued*)

Table 2 (*continued*)

Study	Year	Disease selected	# of outcomes being divided	Data size and source	Sample division	Feature type	ML methods used	Feature selection	# of features being selected	CV	Balancing	Best algorithm	Classification	Calibration	Clinical usefulness	External validation
(34)	2019	CP	2	23 infants with CP; 21 healthy controls	8:2	Genetic features	DL, SVM, RF, LR	No	No	10-fold	No	DL	AUC: 0.976 (95% CI: 0.676–1.00)	No	No	No
(37)	2020	PDDs	2	66 pregnant women with PDDs; 29 healthy controls	No	UtA	LR; RF	Correlation-based and backward greedy stepwise search	4	10-fold	No	RF	AUC: 0.976, PRC: 0.958	No	No	No
(35)	2021	IUGR	2	102 infants with IUGR; 160 healthy controls	No	CTG	SVM	RFE	No	10-fold	No	SVM	Accuracy: 0.93, sensitivity: 0.93, specificity: 0.84	No	No	No
(36)	2020	IUGR	2	60 infants with IUGR; 60 healthy controls	No	CTG	LR, RF, SVM	Fetal heart rate-based encompassing time domain, frequency, and non-linear domains approaches	No	10-fold	No	RF	AUC: 0.911 (95% CI: 0.860–0.961)	No	No	No
(38)	2019	IUGR	2	40 infants with IUGR; 40 healthy controls	No	Metabolomic	SVM	CFS, PLS, LVQ	7	10-fold	No	SVM	AUC: 0.91	No	No	No

AUC, area under the curve; bRF, balanced random forest; CFS, correlation-based feature selection; CI, confidence interval; CNN, convolutional neural network; CP, cerebral palsy; CTG, cardiotocographic; CV, cross-validation; DL, deep learning; DLCRN, deep-learning clinical-radiomics nomogram; DWIC, diffusion-weighted imaging connectome; EEG, electroencephalogram; EEH, extra encephalic health; FBA, fixel-based analysis; GBDT, gradient boosting decision tree; GMM, gaussian mixture model; HIE, hypoxic-ischemic encephalopathy; ICH, intracranial hemorrhage; IUGR, intrauterine growth restriction; KNN, K nearest neighbor; LASSO, least absolute shrinkage and selection operator; LDA, linear discriminant analysis; LR, logistic regression; LSTM, long short-term memory; LVQ, learning vector quantization; ML, machine learning; MLP, multilayer perceptron; MRI, magnetic resonance imaging; NE, neonatal encephalopathy; PA, perinatal asphyxia; PCA, principal component analysis; PDDs, placental dysfunction–related disorders; PLS, partial least square regression; PRC, precision-recall curve; RF, random forest; RFE, recursive feature elimination; SGH, subgaleal hemorrhage; SMOTE, synthetic minority over-sampling technique; SVM, support vector machine; Th, thresholding; UtA, uterine artery; XGBoost, eXtreme Gradient Boosting; Z_{ADC}, apparent diffusion coefficient Z-scores.

the specific problem. Most of the aforementioned studies used only RF models for predictive model development (10,22,29,32), but a single model cannot deal with all NE problems. RF has a number of advantages, including high accuracy, strong generalization ability, and output importance ability. However, it cannot classify small samples or low-dimensional data well, and it has some defects in handling noise problems. Other ML methods also have their own advantages and disadvantages. Thus, in the actual modeling process, according to the "no free lunch" idea, multiple models should be used for development and comparison where possible. For example, Signorin *et al.* (36) used multiple models for modeling and evaluation. The best prediction model should ultimately be selected based on specific needs and then subsequently applied in practice. In addition, in the selection of multiple models, different models have advantages and disadvantages for classification or regression problems, and the applicability of the models should be considered.

Classification, calibration, clinical usefulness, and external validation are all evaluation metrics that measure different dimensions of a model's ultimate applicability. In most cases, researchers report on the discriminatory ability of the model. However, most studies reported only point estimates of AUC values. For example, Guedalia *et al.* (30) reported an AUC1 of 0.88 (95% CI: 0.856–0.904), Zeng *et al.* (32) reported an AUC2 of 0.969; however, we cannot state whether Model 2 is superior to Model 1, as Model 2 is only a point estimate of discrimination, and it was unable to identify the specific problem. However, we can state that Model 1 is an excellent model. Similarly, Bahado-Singh *et al.* (34) reported an AUC of 0.976 (95% CI: 0.676–1.00). However, based on the CI, we can state that the model performed poorly and is not highly suitable for clinical application.

In terms of calibration, it is unfortunate that no study [other than that of Jin *et al.* (15)] reported on model calibration. The calibration of above prediction models was fine; thus, we can assume that their models are highly accurate. As the other studies did not report on model calibration, we cannot evaluate whether the models overestimated or underestimated the risk of disease occurrence.

The decision curve analysis (DCA) method was proposed in 2006 and provides a broad measure of clinical usefulness. Its focus is on the patient benefit at different thresholds during the application of the model. Due to its integration of patient-specific preferences over decision makers, we could not examine the clinical application of a model in a given situation. However, in the field of ML, DCA is rarely used to assess the clinical utility of a model.

The purpose of external validation is to demonstrate the "generalize" ability of a model (i.e., to data sets other than the modeled data). The external validation provides evidence of whether a model is overfitted and assesses its applicability. However, it also places a higher demand on the researcher's workload. A model that has not been externally validated refers to a model that has been applied to clinical data but has not had any real-world applications. Unfortunately, among the included studies, only Raurale *et al.* (24) reported external validation results for their model, which had an accuracy of 0.70 (95% CI: 0.65–0.74) and an internal accuracy of 0.89. However, the accuracy of their model can be considered good to some extent.

In summary, we found that the quality of research on prediction models in most different regions and journals was not adequate. A large number of clinical prediction models cannot be practically applied in practice. Thus, we recommend the use of the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement for clinical diagnosis or prediction model building in the field of ML for NE. The TRIPOD (43) statement was developed by a consortium of clinical prediction model researchers, including statisticians, epidemiologists, methodologists, health care professionals, and journal editors (from the *Annals of Internal Medicine*, *BMJ*, *Journal of Clinical Epidemiology*, and *PLoS Medicine*). The development of predictive models according to the TRIPOD statement provides good insights into the true application of the models to research, and not just the use of ML methods in the field of NE research.

Outcomes associated with NE were classified as dichotomous

The most common disorder in the field of NE is HIE, and most research classifies outcomes using only two data labels (i.e., HIE and no HIE). This classifies the clinical problem as a dichotomous problem, and ignores the categories of mild, moderate, and severe HIE. This division simplifies the differences in HIE brought about by degree by considering only the presence or absence of disease. Such measures are able to improve model performance to some extent. Mooney *et al.* (22) reported an AUC of 0.89 for a typing model built according to the disease versus no disease dichotomy. However, the performance of a multi-categorical model built using the one-to-many strategy showed varying degrees of degradation. This

might be due to the fact that there are certain overlapping features between the different subtypes that restrict the models' ability to differentiate effectively. However, these overlapping features are a manifestation of the complexity of real-world cases, and are the clinical issues that we would prefer to understand in most cases. Focusing on the areas of overlap and differentiating among them effectively was the dilemma facing the study described above.

ML methods are able to distinguish between these overlaps using unsupervised methods such as hierarchical clustering and PCA. This allows the researcher to divide the data into multiple layers for analysis before performing supervised learning. The results of these hierarchies can enhance our knowledge and analysis of the final results. Thus, we are of the view that an in-depth unsupervised learning analysis of the data should be performed prior to supervised learning to deepen the researcher's understanding of the actual problem.

Can ML improve clinical efficiency?

MRI, EEG, and other graphical data are more focused on the diagnosis of diseases related to NE. However, these data are usually obtained when the neonates are already sick. As a result, subsequent interventions lag behind disease progression. The ML model based on graphical data usually performs well from a discrimination perspective. However, from a public health perspective, the cost-effectiveness of such interventions is much lower than that of models that predict the occurrence of NE-related disease using information about the pregnant woman and the birth history of the newborn.

Moreover, some existing studies have reported that postnatal serum measures of specific molecules, such as melatonin, endocannabinoids, and micronutrients like thiamine and beta-hydroxybutyrate, may exert protective effects against NE-related diseases in neonates (44-46). Therefore, incorporating these variables into models could enhance their clinical effectiveness and performance in terms of serum postnatal measures. At the same time, most previous studies have focused on a single dimensional prediction model for a particular disease of brain injury. However, it is difficult to obtain imaging information, an exact diagnosis, or complete biochemical indicators in less developed regions with hierarchical treatment systems. This increases the demand for the applicability of models. Thus, future research needs to develop predictive models for NE that are truly applicable at the primary level in clinical practice.

Data and feature perspective

Data leakage

The following steps are usually employed in the construction of prediction models for most NE diseases: (I) data preprocessing; (II) data division; (III) model training; and (IV) model validation. However, carrying out preprocessing before dividing the data often results in the problem of data leakage. Indirect data leakage refers to leakage that occurs as a result of preprocessing data before data division. Indirect leakage refers to the model prematurely observing the data patterns of the test set (47). Such leakages may significantly affect performance evaluations. Data leakage can occur whenever data preprocessing is performed on both the training set and the test set. Turova *et al.* (29) used the synthetic minority over-sampling technique (SMOTE) for a cross-validation dataset when training a model to deal with a possible data imbalance in cross-validation (48), which also had some data leakage. Data leakage can also occur as a result of normalization and missing value handling, which are commonly used in most studies. Thus, in the above-mentioned model construction process, future research should pay attention to the preprocessing of the data, which should be done after data division to avoid the possibility of data leakage.

Feature engineering

Feature engineering consists of feature construction, extraction, and selection (49). Feature construction refers to the processing of features using normalization, discretization, one-hot coding, metric aggregation, metric transformation, and metric calculation. Feature extraction uses certain methods to effectively reduce the dimensionality of high-dimensional data or imaging data, and to obtain more physically or statistically significant features. Feature selection refers to the use of various statistics or algorithms to filter out a small subset of variables that are more predictive and usually contain filtered, wrapped, and embedded methods.

In the above-mentioned studies, the feature selection processes usually used filtered methods, such as the prior selection of variables using correlation coefficients, and chi-square tests. Filtered methods are computationally efficient, but it does not take into account subsequent learners, and often tends to incorporate redundant features. The wraparound approach wraps feature selection and the learner together, which in turn filters out the subset of features that can improve performance of the learner.

The features incorporated in this way are generally more accurate and result in better model performance, but have high computational complexity, such as that related to RFE used by Turova *et al.* (29) and Pini *et al.* (35).

Embedded methods combine feature selection with ML algorithms that enable the classifier to automatically select features, most commonly typified by the cross-validated least absolute shrinkage and selection operator (LASSO) method used by O'Boyle *et al.* (14). The final model that is built needs to focus on issues such as the economics and practicality of the features to determine their feasibility in clinical practice, such as whether blood gas analysis data should be incorporated into the model as mentioned by Mooney *et al.* (22). Additionally, there are differences in the features screened by different feature selection methods. Their performance on experimental data may be good, but eventually, the suitability of building models based on certain features needs to be verified on external validation datasets.

Data collection and the clinical dilemma

Data acquisition is the most critical issue not only in the research process but also in the future clinical practice. A possible problem related to the use of imaging data such as MRI data in the neonatal field is that neonates need to be sedated to undergo MRI scans. This places an extremely high demand on data acquisition. However, appropriate protocols are currently being implemented at various centers that allow neonates to undergo MRI scans without the need for sedation. These methods are considered both safe and effective.

Today, primary care facilities are unable to accommodate the acquisition of imaging data, and individualized care is needed to diagnose NE-related disorders using clinically accessible data. Thus, early warning prediction models for NE could be established using easily available data such as maternal factors or ultrasound. The risk of NE can be indicated early, and appropriate interventions can be implemented before or after the birth of the newborn. The aim is to reduce the risk of future neurodevelopmental disorders in newborns and the socioeconomic burden placed on families, and to increase the healthy life expectancy of newborns.

Conclusions

To date, most studies on the application of ML in NE have not comprehensively considered issues related to the experimental design, data processing, model building, and evaluation. The aforementioned issues not only apply to the selection of ML methods, but also apply to the standardized modeling process, data leakage, feature engineering, and model evaluation. Regarding the final issue of translating research into practice, to date, few studies have successfully implemented models in real clinical settings. The non-standardized model development process partly explains the poor utility of the final models in the existing studies. In this article, we examined many studies in which ML was applied to NE, and abstractly explored the limitations and challenges encountered by most of them. Some possible solutions were proposed. In future research and applications, we hope that diagnostic prediction models with applications to primary or hierarchical diagnostic and treatment systems will emerge. These models could provide effective decision-making tools for clinicians, and thus improve the healthy life span of newborns.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the Narrative Review reporting checklist. Available at https://tp.amegroups.com/article/view/10.21037/tp-24-425/rc

Peer Review File: Available at https://tp.amegroups.com/article/view/10.21037/tp-24-425/prf

Funding: None.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tp.amegroups.com/article/view/10.21037/tp-24-425/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. Loverro G, De Cosmo L, Loverro M, et al. Neonatal Encephalopathy. In: Malvasi A, Tinelli A, Di Renzo G, editors. Management and Therapy of Late Pregnancy Complications. Cham: Springer; 2017:359-67.
2. Executive summary: Neonatal encephalopathy and neurologic outcome, second edition. Report of the American College of Obstetricians and Gynecologists' Task Force on Neonatal Encephalopathy. Obstet Gynecol 2014;123:896-901.
3. De Angelis LC, Brigati G, Polleri G, et al. Neonatal Hypoglycemia and Brain Vulnerability. Front Endocrinol (Lausanne) 2021;12:634305.
4. Newton CR. Global Burden of Pediatric Neurological Disorders. Semin Pediatr Neurol 2018;27:10-5.
5. Lee AC, Kozuki N, Blencowe H, et al. Intrapartum-related neonatal encephalopathy incidence and impairment at regional and global levels for 2010 with trends from 1990. Pediatr Res 2013;74 Suppl 1:50-72.
6. Dammann O, Ferriero D, Gressens P. Neonatal encephalopathy or hypoxic-ischemic encephalopathy? Appropriate terminology matters. Pediatr Res 2011;70:1-2.
7. Martinello K, Hart AR, Yap S, et al. Management and investigation of neonatal encephalopathy: 2017 update. Arch Dis Child Fetal Neonatal Ed 2017;102:F346-58.
8. Schump EA. Neonatal Encephalopathy: Current Management and Future Trends. Crit Care Nurs Clin North Am 2018;30:509-21.
9. Navarro X, Porée F, Kuchenbuch M, et al. Multi-feature classifiers for burst detection in single EEG channels from preterm infants. J Neural Eng 2017;14:046015.
10. Jeong JW, Lee MH, Fernandes N, et al. Neonatal encephalopathy prediction of poor outcome with diffusion-weighted imaging connectome and fixel-based analysis. Pediatr Res 2022;91:1505-15.
11. Kim KY, Nowrangi R, McGehee A, et al. Assessment of germinal matrix hemorrhage on head ultrasound with deep learning algorithms. Pediatr Radiol 2022;52:533-8.
12. Deng H, Zhou Y, Wang L, et al. Ensemble learning for the early prediction of neonatal jaundice with genetic features. BMC Med Inform Decis Mak 2021;21:338.
13. Song C, Jiang ZQ, Liu D, et al. Application and research progress of machine learning in the diagnosis and treatment of neurodevelopmental disorders in children. Front Psychiatry 2022;13:960672.
14. O'Boyle DS, Dunn WB, O'Neill D, et al. Improvement in the Prediction of Neonatal Hypoxic-Ischemic Encephalopathy with the Integration of Umbilical Cord Metabolites and Current Clinical Makers. J Pediatr 2021;229:175-181.e1.
15. Jin MC, Parker JJ, Rodrigues AJ, et al. Development of an integrated risk scale for prediction of shunt placement after neonatal intraventricular hemorrhage. J Neurosurg Pediatr 2022;29:444-53.
16. Zwanenburg A, Andriessen P, Jellema RK, et al. Using trend templates in a neonatal seizure algorithm improves detection of short seizures in a foetal ovine model. Physiol Meas 2015;36:369-84.
17. Escobar GJ, Soltesz L, Schuler A, et al. Prediction of obstetrical and fetal complications using automated electronic health record data. Am J Obstet Gynecol 2021;224:137-147.e7.
18. Gupta C, Chandrashekar P, Jin T, et al. Bringing machine learning to research on intellectual and developmental disabilities: taking inspiration from neurological diseases. J Neurodev Disord 2022;14:28.
19. Dutton DM, Conroy GV. A review of machine learning. Knowl Eng Rev 1997;12:341-67.
20. Tataranno ML, Vijlbrief DC, Dudink J, et al. Precision Medicine in Neonates: A Tailored Approach to Neonatal Brain Injury. Front Pediatr 2021;9:634092.
21. Samanta D. Recent Advances in the Diagnosis and Treatment of Neonatal Seizures. Neuropediatrics 2021;52:73-83.
22. Mooney C, O'Boyle D, Finder M, et al. Predictive modelling of hypoxic ischaemic encephalopathy risk following perinatal asphyxia. Heliyon 2021;7:e07411.
23. Pavel AM, O'Toole JM, Proietti J, et al. Machine learning for the early prediction of infants with electrographic seizures in neonatal hypoxic-ischemic encephalopathy. Epilepsia 2023;64:456-68.
24. Raurale SA, Boylan GB, Mathieson SR, et al. Grading hypoxic-ischemic encephalopathy in neonatal EEG with convolutional neural networks and quadratic time-frequency distributions. J Neural Eng 2021;18:046007.
25. Weiss RJ, Bates SV, Song Y, et al. Mining multi-site clinical data to develop machine learning MRI biomarkers: application to neonatal hypoxic ischemic encephalopathy. J

Transl Med 2019;17:385.
26. Zheng Q, Martin-Saavedra JS, Saade-Lemus S, et al. Cerebral Pulsed Arterial Spin Labeling Perfusion Weighted Imaging Predicts Language and Motor Outcomes in Neonatal Hypoxic-Ischemic Encephalopathy. Front Pediatr 2020;8:576489.
27. Tian T, Gan T, Chen J, et al. Graphic Intelligent Diagnosis of Hypoxic-Ischemic Encephalopathy Using MRI-Based Deep Learning Model. Neonatology 2023;120:441-9.
28. Lew CO, Calabrese E, Chen JV, et al. Artificial Intelligence Outcome Prediction in Neonates with Encephalopathy (AI-OPiNE). Radiol Artif Intell 2024;6:e240076.
29. Turova V, Sidorenko I, Eckardt L, et al. Machine learning models for identifying preterm infants at risk of cerebral hemorrhage. PLoS One 2020;15:e0227419.
30. Guedalia J, Lipschuetz M, Daoud-Sabag L, et al. Prediction of neonatal subgaleal hemorrhage using first stage of labor data: A machine-learning based model. J Gynecol Obstet Hum Reprod 2022;51:102320.
31. Zhang X, Zeng S, Cheng G, et al. Clinical Manifestations of Neonatal Hyperbilirubinemia Are Related to Alterations in the Gut Microbiota. Children (Basel) 2022;9:764.
32. Zeng S, Wang Z, Zhang P, et al. Machine learning approach identifies meconium metabolites as potential biomarkers of neonatal hyperbilirubinemia. Comput Struct Biotechnol J 2022;20:1778-84.
33. Chou JH. Predictive Models for Neonatal Follow-Up Serum Bilirubin: Model Development and Validation. JMIR Med Inform 2020;8:e21222.
34. Bahado-Singh RO, Vishweswaraiah S, Aydas B, et al. Deep Learning/Artificial Intelligence and Blood-Based DNA Epigenomic Prediction of Cerebral Palsy. Int J Mol Sci 2019;20:2075.
35. Pini N, Lucchini M, Esposito G, et al. A Machine Learning Approach to Monitor the Emergence of Late Intrauterine Growth Restriction. Front Artif Intell 2021;4:622616.
36. Signorini MG, Pini N, Malovini A, et al. Integrating machine learning techniques and physiology based heart rate features for antepartum fetal monitoring. Comput Methods Programs Biomed 2020;185:105015.
37. Sufriyana H, Wu YW, Su EC. Prediction of Preeclampsia and Intrauterine Growth Restriction: Development of Machine Learning Models on a Prospective Cohort. JMIR Med Inform 2020;8:e15411.
38. Bahado-Singh RO, Yilmaz A, Bisgin H, et al. Artificial intelligence and the analysis of multi-platform metabolomics data for the detection of intrauterine growth restriction. PLoS One 2019;14:e0214121.
39. Lu J, Gong P, Ye J, et al. A survey on machine learning from few samples. Pattern Recognit 2023;139109480.
40. Karamizadeh S, Abdullah SM, Halimi M, et al. Advantage and drawback of support vector machine functionality. 2014 International Conference on Computer, Communications, and Control Technology (I4CT), Langkawi, Malaysia, 2014:63-5.
41. Lin WC, Tsai CF. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev 2020;53:1487-509.
42. Wolpert DH. The Supervised Learning No-Free-Lunch Theorems. In: Roy R, Köppen M, Ovaska S, et al. editors. Soft Computing and Industry. London: Springer; 2002:25-42.
43. Collins GS, Reitsma JB, Altman DG, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. Br J Surg 2015;102:148-58.
44. Sechi G, Sechi MM. New Therapeutic Paradigms in Neonatal Hypoxic-Ischemic Encephalopathy. ACS Chem Neurosci 2023;14:1004-6.
45. Sechi GP, Bardanzellu F, Pintus MC, et al. Thiamine as a Possible Neuroprotective Strategy in Neonatal Hypoxic-Ischemic Encephalopathy. Antioxidants (Basel) 2021;11:42.
46. Hassell KJ, Ezzati M, Alonso-Alconada D, et al. New horizons for newborn brain protection: enhancing endogenous neuroprotection. Arch Dis Child Fetal Neonatal Ed 2015;100:F541-52.
47. Samala RK, Chan HP, Hadjiiski L, et al. Hazards of data leakage in machine learning: a study on classification of breast cancer using deep neural networks, Medical Imaging 2020: Computer-Aided Diagnosis, SPIE, 2020:279-84.
48. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-57.
49. Nargesian F, Samulowitz H, Khurana U, et al. Learning Feature Engineering for Classification. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017:2529-35.

Cite this article as: Huang YF, Jiang ZQ, Feng L, Song C. Current progress and future prospects of machine learning in the diagnosis of neonatal encephalopathy: a narrative review. Transl Pediatr 2025;14(4):728-739. doi: 10.21037/tp-24-425