Psychiatry Research 320 (2023) 115050

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Early identification of autism spectrum disorder by multi-instrument fusion: A clinically applicable machine learning approach

Qiuhong Wei [a,1], Xueli Xu [b,1], Ximing Xu [c,*], Qian Cheng [a,*]

[a] Department of Children's Healthcare, National Clinical Research Center for Child Health and Disorders, Ministry of Education Key Laboratory of Child Development and Disorders, China International Science and Technology Cooperation Base of Child development and Critical Disorders, Chongqing Key Laboratory of Childhood Nutrition and Health, Children's Hospital of Chongqing Medical University, No 136. Zhongshan 2nd Rd, Yuzhong District, Chongqing, China
[b] Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, China
[c] Big Data Center for Children's Medical Care, Children's Hospital of Chongqing Medical University, No 136. Zhongshan 2nd Rd, Yuzhong District, Chongqing 400014, China

ARTICLE INFO

Keywords:
Neurodevelopmental disorder
Developmental language disorder
Global developmental delay
Behavior scale

ABSTRACT

Autism spectrum disorder (ASD), developmental language disorder (DLD), and global developmental delay (GDD) are common neurodevelopmental disorders in early childhood; however, the differential diagnosis of these disorders is difficult because of overlapping symptoms. Drawing on a cohort of 2004 children with ASD, DLD, or GDD, this study developed machine learning classifiers using decision trees, support vector machines, eXtreme gradient boosting (XGB), logistic regression, and neural networks by combining several easily accessible behavioral and developmental assessment instruments. The best-performing XGB model was further simplified into a two-stage decision model (TS-DM) to achieve better interpretability. Model performance was tested and compared with that of 12 pediatricians on an external dataset of 60 children. The accuracies of the resident pediatricians, senior pediatricians, TS-DM, and XGB were 53.3%, 66.7%, 75.0%, and 78.3%, respectively. Machine learning has the potential to identify these three neurodevelopmental disorders by integrating information from multiple instruments and thereby may increase our understanding of the roles of different behavioral and developmental characteristics in the different diagnoses.

1. Introduction

Autism Spectrum Disorder (ASD) is a heterogeneous neurodevelopmental disorder characterized by impaired social communication, repetitive behaviors, highly restricted interests, and/or abnormal sensory reactions. As recently reported, the prevalence of ASD was 2.3% in the United States (Maenner et al., 2021) and 0.7% in China (Zhou et al., 2020). Notably, the prevalence of ASD has gradually increased and has become a serious global public health issue (Collaborators, 2022; Lyall et al., 2017).

Language delay is the chief complaint of most caregivers of children with ASD in early childhood. However, developmental language disorder (DLD) is also characterized by persistent deficits in the acquisition, understanding, production, or use of language (Bishop et al., 2017), while global developmental delay (GDD) is defined as a significant delay in two or more developmental domains (speech and language, gross or fine motor skills, cognition, personal-social, and activities of daily living)(Mithyantha et al., 2017). Therefore, the overlapping symptoms of the three disorders may confound the diagnosis.

In clinical practice, pediatricians formulate a differential diagnosis of the three disorders based on medical history, clinical observation, and results of assessment instruments (Genovese and Butler, 2020). During a consultation with a child who manifests language delay, pediatricians usually start with a language measure, such as the Early Language Milestone Scale (ELMS), to assess the child's language ability (Larson, 2016). Then the autism assessment instruments (e.g., Modified Checklist for Autism in Toddlers (M-CHAT), Autism Behavior Checklist (ABC), and Autism Diagnostic Observation Schedule) are used to assess autistic symptoms to identify ASD (Lord et al., 2018; Toh et al., 2018). Finally, a developmental scale (e.g., Gesell Developmental Scale (GDS), Mullen Scales of Early Learning, and Bayley Scales) is used to assess cognitive level so as to establish GDD (Del et al., 2021; Farmer et al., 2016; Liu

* Corresponding authors.
E-mail addresses: ximing@hospital.cqmu.edu.cn (X. Xu), chengqq@cqmu.edu.cn (Q. Cheng).
[1] Qiuhong Wei and Xueli Xu are co-first authors of the article.

https://doi.org/10.1016/j.psychres.2023.115050
Received 3 June 2022; Received in revised form 30 December 2022; Accepted 5 January 2023
Available online 10 January 2023
0165-1781/© 2023 Elsevier B.V. All rights reserved.

et al., 2018). However, diagnostic instruments, such as the Autism Diagnostic Observation Schedule and the Autism Diagnostic Interview-Revised, require extensive and costly training, limiting their widespread use, especially in primary care and developing countries (Bauer et al., 2022; Clark et al., 2019). Pediatricians in regions where medical resources are scarce commonly use easy-to-acquire and low-cost instruments (e.g., M-CHAT, ABC, GDS, and ELMS) (Larson, 2016; Marlow et al., 2019). In addition, for less experienced resident pediatricians, it is difficult to synthesize results from several assessments to form a reliable diagnosis. Therefore, we aimed to develop a machine learning model that utilizes these easy-to-acquire instruments as a decision support system to assist in the early identification of children with ASD, DLD, or GDD.

contains info for text and speech based methods

In recent years, machine learning has been widely used to build clinical decision support systems with applications in aiding medical diagnosis, reducing missed diagnoses and misdiagnoses, and improving diagnostic efficiency (Jacob et al., 2019; Liang et al., 2019; Ngiam and Khor, 2019). Machine learning combined with behavioral instruments has been widely used for identifying ASD (Cavus et al., 2021). Most of the established classifiers have used autism assessment instruments (e.g., M-CHAT, Autism Diagnostic Interview-Revised, and Social Responsiveness Scale) to distinguish children with ASD from typically developing children (Duda et al., 2014; Goel et al., 2020; Shahamiri et al., 2022; Wall et al., 2012). Few studies have utilized information from language and cognitive assessment instruments, which play an essential role in the differential diagnosis of ASD and other neurodevelopmental disorders.

To address these knowledge gaps, we first developed a machine learning approach using multiple easy-to-access instruments as a decision support system to assist in the early identification of children with ASD, DLD, or GDD. Second, we further simplified the machine learning model and visualized its classification process to achieve better clinical interpretability, which could help less-experienced pediatricians improve their diagnostic efficiency. Finally, we translated the developed model into a clinician-friendly web application that was easier to use while providing real-time decision support.

2. Methods

2.1. Ethics statement

This study was approved by the Ethics Committee of the Children's Hospital of Chongqing Medical University and informed consent was waived (File No. 2,021,052).

2.2. Study population

This cross-sectional study was conducted from January 2019 to December 2021 at the National Clinical Research Center for Child Health and Disorders, Chongqing, China. Data for this study were collected from the Child Healthcare Database. The following inclusion criteria were applied: children with ASD, DLD, or GDD were diagnosed by developmental-behavioral pediatricians based on the Diagnostic and Statistical Manual of Mental Disorders (Fifth Edition) and diagnostic assessment instruments; all children were assessed using the Gesell Developmental Scale (GDS), the Early Language Milestone Scale (ELMS), the Modified Checklist for Autism in Toddlers (M-CHAT), and the Autism Behavior Checklist (ABC). Patients were excluded if they were diagnosed with hearing impairment or any neurological disease (e.g., epilepsy, Wilson's disease, autoimmune encephalitis) or if their assessment data were missing.

2.3. Data preparation

A total of 14 features from 2004 children with ASD, DLD, or GDD were used for model development. These 14 features included two demographic characteristics (age and sex) and 12 summary metrics from the four assessment instruments: five developmental quotients from the GDS, four developmental quotients from the ELMS, the ABC total score, and two features from the M-CHAT.

In particular, the GDS consists of tasks that measure gross motor, fine motor, language, personal-social, and adaptive domains. The assessment results from the GDS were summarized as five development quotients in the five domains. The ELMS consists of evaluations of four language domains, namely expressive language, receptive language, visual-related language, and overall language ability. The results from the ELMS were reported as developmental age, which could be converted to a development quotient by dividing developmental age by chronological age and multiplying by 100. The ABC is a 57-item questionnaire that measures the severity of autistic symptoms: abnormal language, abnormal body and object use, difficulties relating to others, sensory issues, and social and self-help difficulties. The result of the ABC was represented as a total score. The M-CHAT is a parent-reported 23-item (including six critical items) questionnaire used to screen for ASD. A positive screen was indicated by failing either (1) two or more of the six critical items or (2) three or more of the 23 items. The results of the M-CHAT were represented as the number of failed items and failed critical items.

2.4. Development of machine learning models

A summary of the study procedure is presented in Fig. 1. Data for the model development were collected from January 2019 to October 2021. Following this, 60 newly diagnosed patients were enrolled from November 2021 to December 2021 as the test dataset. For model development, we randomly selected 60 children (20 per group) with the three disorders (ASD, GDD, and DLD) for validation and the remainder for training. This process was repeated 20 times, and the average and standard deviation of the predictive results were recorded.

Five well-established and popular machine learning algorithms, which have been applied to build machine learning models for identifying children with ASD, were chosen to construct classification models to identify children with ASD, DLD, or GDD (Cavus et al., 2021; Deo, 2015; Fan et al., 2021; Zhang et al., 2021). The five algorithms included Decision Tree (DT) (Song and Lu, 2015), Support Vector Machine (SVM) (Cortes and Vapnik, 1995), L1-regularized Logistic Regression (LR) (Tibshirani, 1996), eXtreme Gradient Boosting (XGB) (Chen and Guestrin, 2016), and Neural Network (NN) (Hinton and Salakhutdinov, 2006). These algorithms were programmed using Python 3.6.7, using a grid search strategy with cross-validation to determine the optimal parameter settings. The detailed parameter settings are shown in the Supplementary Material.

2.5. A two-stage decision model

The XGB is a complex multi-tree model for classifying the three neurodevelopmental disorders, and its decision process cannot be clearly visualized and is difficult to explain. In this study, we simplified the structure of the XGB model into a Two-stage decision model (TS-DM) to make the decision process easier to interpret. Following Yan et al. (2020), we first identified a tree in the XGB model that contributed the most to the classification of the DLD and the combination of ASD and GDD. A single-tree model was then constructed with a clinically operable decision process to classify DLD and other disorders. Similarly, in the second stage, a single-tree model was built to classify children with ASD or GDD. The detailed model development process is described in the Supplementary Material.

2.6. Pediatrician-machine competition

According to the guideline of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

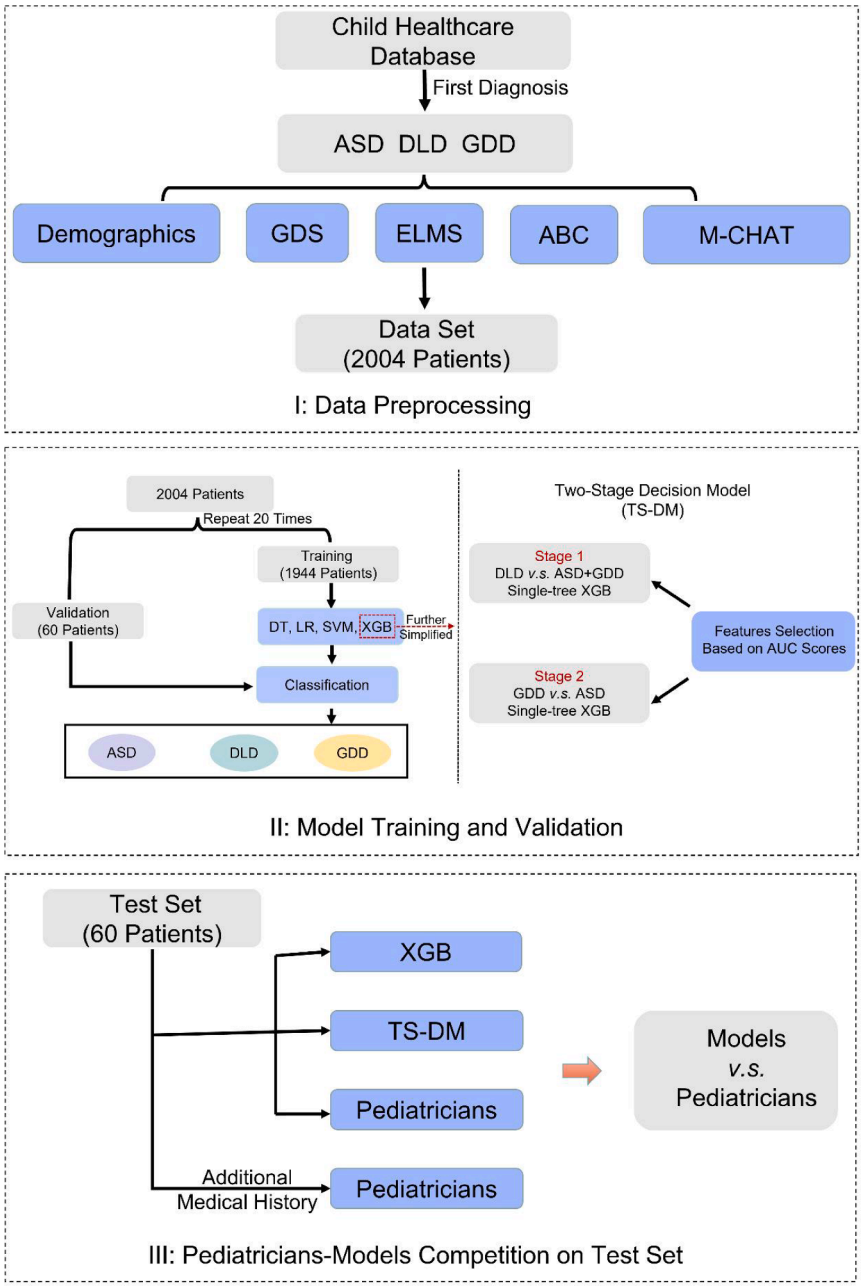

Fig. 1. Diagram of the study pipeline. (I) Data preprocessing; (II) Flowchart of model training and validation; (III) Pediatricians-models competition in the test set.
DLD: Developmental Language Disorder; GDD: Global Developmental Delay; ASD: Autism Spectrum Disorder; M-CHAT: Modified Checklist for Autism in Toddlers; ABC: Autism Behavior Checklist; GDS: Gesell Developmental Scale; ELMS: Early Language Milestone Scale; XGB: eXtreme Gradient Boosting; AUC: Area Under the Curve.

(TRIPOD) statement (Collins et al., 2015), the classification performance of the developed machine learning model should be further evaluated on an external dataset (also known as the test set). To better understand the value of the model in clinical applications, we conducted a pediatricians-machine competition on the test set (60 patients) in addition to evaluating the performance of the model itself. If the model performs worse than the pediatricians, the model will then have very limited or no value in clinical practice.

For the pediatricians-machine competition, twelve pediatricians, including six resident pediatricians with less than three years of clinical practice experience and six senior pediatricians with more than ten years of experience, were recruited. The pediatricians who participated in the pediatricians-machine competition were different from those who diagnosed these children initially. We randomly assigned five patients to each of the 12 pediatricians. The diagnostic accuracies of resident pediatricians, senior pediatricians, and machine learning models were compared in the following two scenarios. First, the pediatricians were provided with electronic reports of assessment instruments, which are the same information used in the machine learning models, including two demographic characteristics (age and sex) and 12 summary metrics from the four assessment instruments (denoted by "Pediatricians" in Fig. 3). We then provided the pediatricians with additional electronic medical history for each child, including the chief complaint, present illness, growth and developmental history, family history, behavioral observation, and physical examination, to allow them to make a diagnosis again (denoted by "Pediatricians^{h}" in Fig. 3). Under each scenario, the pediatricians selected the most likely diagnosis among ASD, GGD, and DLD. All pediatricians diagnosed the patients independently and submitted their decisions within three days.

2.7. Model evaluation metrics

To evaluate the performance of the developed machine learning models in classifying the three neurodevelopmental disorders, we

considered two commonly used metrics, namely accuracy and recall. Accuracy was the percentage of correct classifications by the model. Recall, also known as sensitivity, measured the proportion of positives correctly classified. They were calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \text{and} \quad \text{Recall} = \frac{TP}{TP + FN},$$

where TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively. The predicted class was assigned using a default decision threshold of 0.5 (Liu et al., 2019; Steyerberg, 2019).

In this study, we used the area under the receiver operating characteristic curve to select the important features for the two-stage decision model. The curve was created by plotting the true positive rate (TPR) against the false positive rate (FPR) on the validation dataset. The TPR and FPR were calculated as follows:

$$\text{TPR} = \frac{TP}{TN+FP} \text{and} \quad \text{FPR} = \frac{FN}{TP+FN}.$$

2.7. Statistical analysis

All statistical analyses were conducted using R (v.4.0.2) software. Continuous variables were described using the median and interquartile range. Categorical variables were expressed as frequencies (percentages). Kruskal-Wallis and Chi-square tests were used to compare differences in continuous variables and categorical variables, respectively, regarding behavioral and developmental characteristics among the three groups. Cohen's kappa statistic was used to test the inter-group diagnostic agreement of the machine learning models and pediatricians. Principal Component Analysis was used to visualize the data in the lower dimension using the prcomp function in the stats package. Statistical significance was set at $P < 0.05$.

3. Results

3.1. Demographic, developmental, and behavioral characteristics of study subjects

The ASD group included 1543 patients (1291 males and 252 females) with a median age of 29.52 months (interquartile range, 25.54 to 35.06). The DLD group included 201 patients (170 males and 31 females) with a median age of 27.03 months (interquartile range, 24.10 to 30.10). The GDD group included 260 patients (221 males and 39 females) with a median age of 32.06 months (interquartile range, 27.74 to 36.32). As shown in Table 1, statistically significant differences were found among these three groups in autistic symptoms (the M-CHAT and the ABC), developmental level (the GDS), and language ability (the ELMS). The Principal Component Analysis plots of the first two principal components in Supplementary Fig. 1 show that these three groups were not clearly separated.

3.2. Feature importance and performances of the two-stage decision model

We noted that the accuracy of all models in identifying DLDs outperformed that of the ASDs and GDDs (Supplementary Table 5). Next, this study combined GDD and ASD into one group (GSD) to develop a two-stage decision model (TS-DM). The ranking features for the two stages through XGB, based on the Shapley values, are shown in Supplementary Tables 1–2. In the first stage, the top four features (adaptive, personal-social, fine motor, and gross motor) were used for the final classification (Stage one in Table 2), and the accuracies for DLD and GSD were 85.50±9.50% and 93.24±4.19%, respectively. In the second stage, the most important feature was the M-CHAT-critical items (Stage two in Table 2), and the accuracies for ASD and GDD were 75.00±9.35% and 76.75±7.63%, respectively. The distributions of these important features in these three groups are shown in Supplementary Fig. 2.

Table 1
Summary of behavior and developmental characteristics from multiple instruments.

Characteristics	ASD ($n = 1543$)	DLD ($n = 201$)	GDD ($n = 260$)	P
M-CHAT, n (%)				
Screening positive	1400(90.7)	133(66.2)	195(75.0)	<0.01
Screening negative	143(9.3)	68(33.8)	65(25.0)	<0.01
ABC, median (Q1, Q3)				
ABC total scores	56(44, 68)	42(30, 58)	46(33, 59)	<0.01
GDS, median (Q1, Q3)				
Gross motor	73(64, 83)	89(83, 95)	74(67, 83)	<0.01
Fine motor	71(58, 86)	94(86, 104)	75(68, 88)	<0.01
Language	41(30, 53)	62(55, 69)	49(39, 58)	<0.01
Adaptive	62(49, 74)	85(79, 91)	67(59, 73)	<0.01
Personal-social	56(48, 66)	73(67, 79)	60(53, 67)	<0.01
ELMS, median (Q1, Q3)				
Expressive language	39(32, 48)	49(40, 58)	42(33, 51)	<0.01
Receptive language	27(22, 37)	46(35, 57)	38(29, 49)	<0.01
Visual related language	36(26, 45)	50(44, 57)	44(37, 50)	<0.01
Overall language ability	36(28, 45)	51(43, 57)	42(36, 49)	<0.01

Q1: lower quartile; Q3: upper quartile; M-CHAT: Modified Checklist for Autism in Toddlers; ABC: Autism Behavior Checklist; GDS: Gesell Developmental Scale; ELMS: Early Language Milestone Scale; ASD: Autism Spectrum Disorder; DLD: Developmental Language Disorder; GDD: Global Developmental Delay. The p-values were calculated with Kruskal-Wallis tests for continuous variables and Chi-square test for categorical variables.

Table 2
Feature selection for TS-DM based on AUC scores of the validation set with 20 random replications.

TS-DM	AUC (mean±sd)	Features
Stage one	0.855±0.048	Adaptive
	0.868±0.039	Adaptive + personal-social
	0.878±0.038	Adaptive + personal-social + fine motor
	0.908±0.027	Adaptive + personal-social + fine motor + gross motor
	0.900±0.034	Adaptive + personal-social + fine motor + gross motor + language
Stage two	0.727±0.053	M-CHAT critical items
	0.727±0.049	M-CHAT critical items + visual related language

TS-DM: two-stage decision model; AUC: area under the receiver operator characteristic curve; sd: standard deviation; M-CHAT: Modified Checklist for Autism in Toddlers; Adaptive: Development Quotient(DQ) of adaptive domain in Gesell Developmental Scale(GDS); personal-social: DQ of personal-social domain in the GDS; fine motor: DQ of fine motor domain in the GDS; gross motor: DQ of gross motor domain in the GDS; language: DQ of language domain in the GDS; M-CHAT critical items: number of failed critical items in M-CHAT; visual related language: DQ of visual related language domain in Early Language Milestone Scale.

3.3. Results of pediatricians-machine competition

A total of 60 children (51 males and 9 females), with a median age of 30.08 months (interquartile range, 26.87 to 35.50), were included in this study as a test set. The decision process of the TS-DM is shown in Fig. 2. The performances of the XGB and TS-DM models, senior pediatricians, and resident pediatricians are presented in Table 3, their classification labels for each of the 60 patients are shown in Fig. 3, and the confusion matrices of classification are displayed in Supplementary Fig. 3. The performances of the developed XGB and TS-DM models were comparable to that of senior pediatricians with access to additional medical history and superior to that of resident pediatricians. The Cohen's kappa values for agreement between original diagnosis and the classification results from XGB, TS DM, and pediatricians with and without using additional medical history were 0.7, 0.6, 0.5 (with

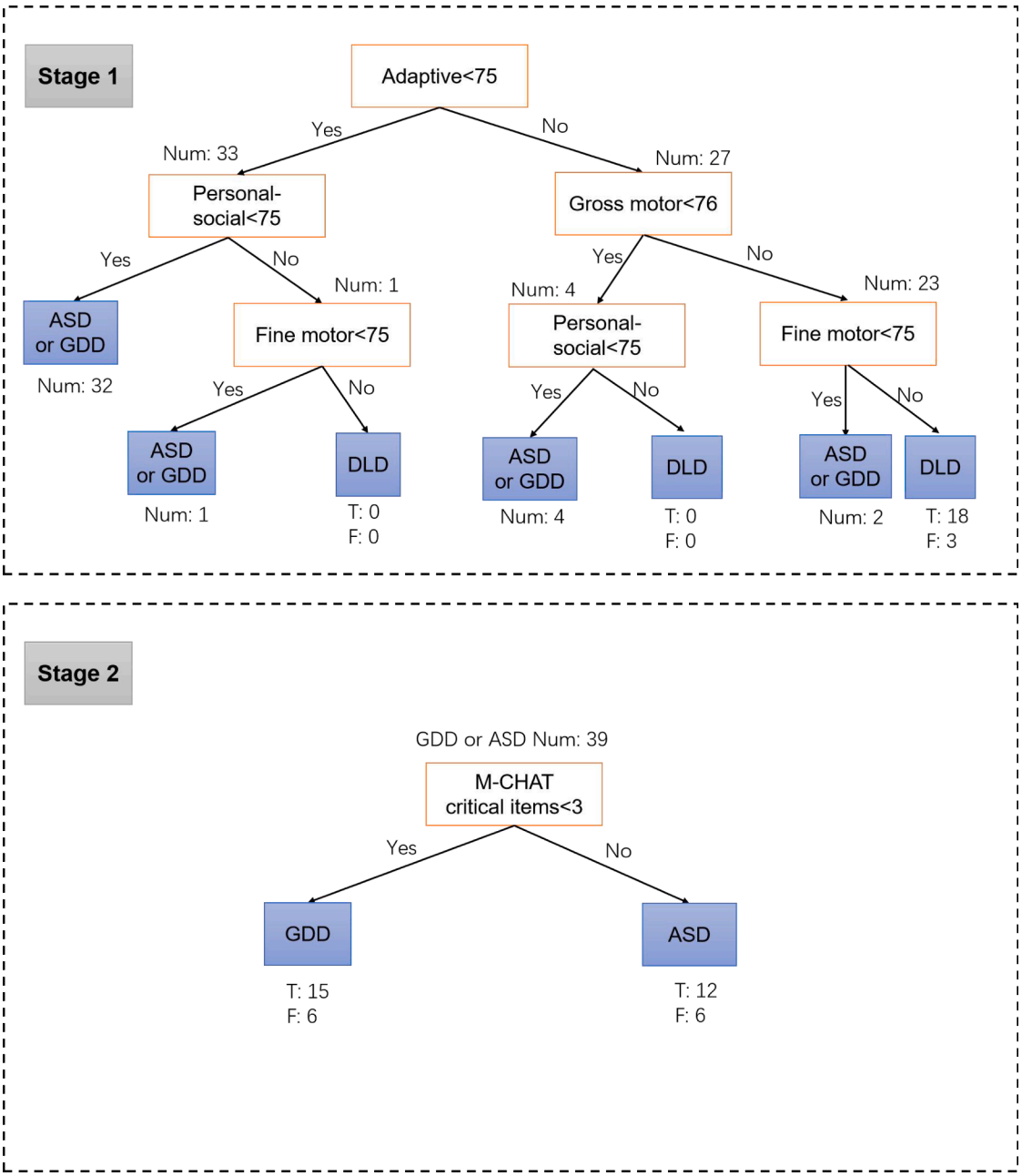

Fig. 2. Illustration of the two-stage decision model on the test dataset of 60 patients, with associated features and thresholds. DLD: Developmental Language Disorder;
GDD: Global Developmental Delay; ASD: Autism Spectrum Disorder. Num: the number of patients in a group; T: the number of correctly classified patients; F: the number of misclassified patients.

additional medical history - senior pediatricians: 0.6, resident pediatricians: 0.4), and 0.4 (without additional medical history - senior pediatricians: 0.5, resident pediatricians: 0.3), respectively (all $P < 0.05$).

4. Discussion

In this study, we developed machine-learning models based on multi-instrument fusion to classify three common neurodevelopmental disorders in early childhood. The machine learning models achieved satisfactory classification accuracy, comparable to that of senior pediatricians. The proposed TS-DM model visualizes the classification process, which could increase our understanding of the roles of different behavioral and developmental characteristics in the differential diagnosis.

To our knowledge, this is the first attempt to apply machine learning to distinguish between three neurodevelopmental disorders by combining developmental and behavioral assessment instruments. Distinguishing children with ASD and typically developing children is relatively easy because of the marked discrepancies in their behaviors. Previous studies have proposed models with high accuracy for classifying children with ASD and typically developing children (Goel et al., 2020; Mujeeb and Monica, 2021). However, in clinical practice, pediatricians are more likely to be confused when distinguishing ASD from DLD and GDD (Hus and Segal, 2021; Miller et al., 2019) rather than identifying children with ASD and typically developing children. In this study, the developed machine learning models achieved a satisfactory classification accuracy, and thus demonstrate the potential to assist pediatricians in the complex differential diagnosis of these three

Table 3
Performances (%) of machine learning models and pediatricians on the test set.

Decision maker	Cases	Accuracy (%)	R for ASD (%)	R for DLD (%)	R for GDD (%)
Pediatrician	60	60.0	70.0	60.0	50.0
Pediatrician[h]		66.7	**75.0**	65.0	60.0
XGB		**78.3**	65.0	**100.0**	70.0
TS-DM		75.0	60.0	90.0	**75.0**
R-Pediatrician	30	53.3	69.2	44.4	37.5
R-Pediatrician[h]		60.0	**76.9**	55.5	37.5
XGB		**80.0**	69.2	**100**	75
TS-DM		73.3	61.5	77.8	**87.5**
S-pediatrician	30	66.7	**71.4**	72.7	58.3
S-pediatrician[h]		73.3	**71.4**	72.7	**75.0**
XGB		**76.7**	57.1	**100.0**	66.7
TS-DM		**76.7**	57.1	**100.0**	66.7

R: recall; h: provided with additional information on medical history; R-Pediatrician: resident pediatrician; S-Pediatrician: senior pediatrician; ASD: Autism Spectrum Disorder; DLD: Developmental Language Disorder; GDD: Global Developmental Delay; XGB: extreme gradient boosting; TS-DM: two-stage decision model. The best values of performance metrics are highlighted in bold.

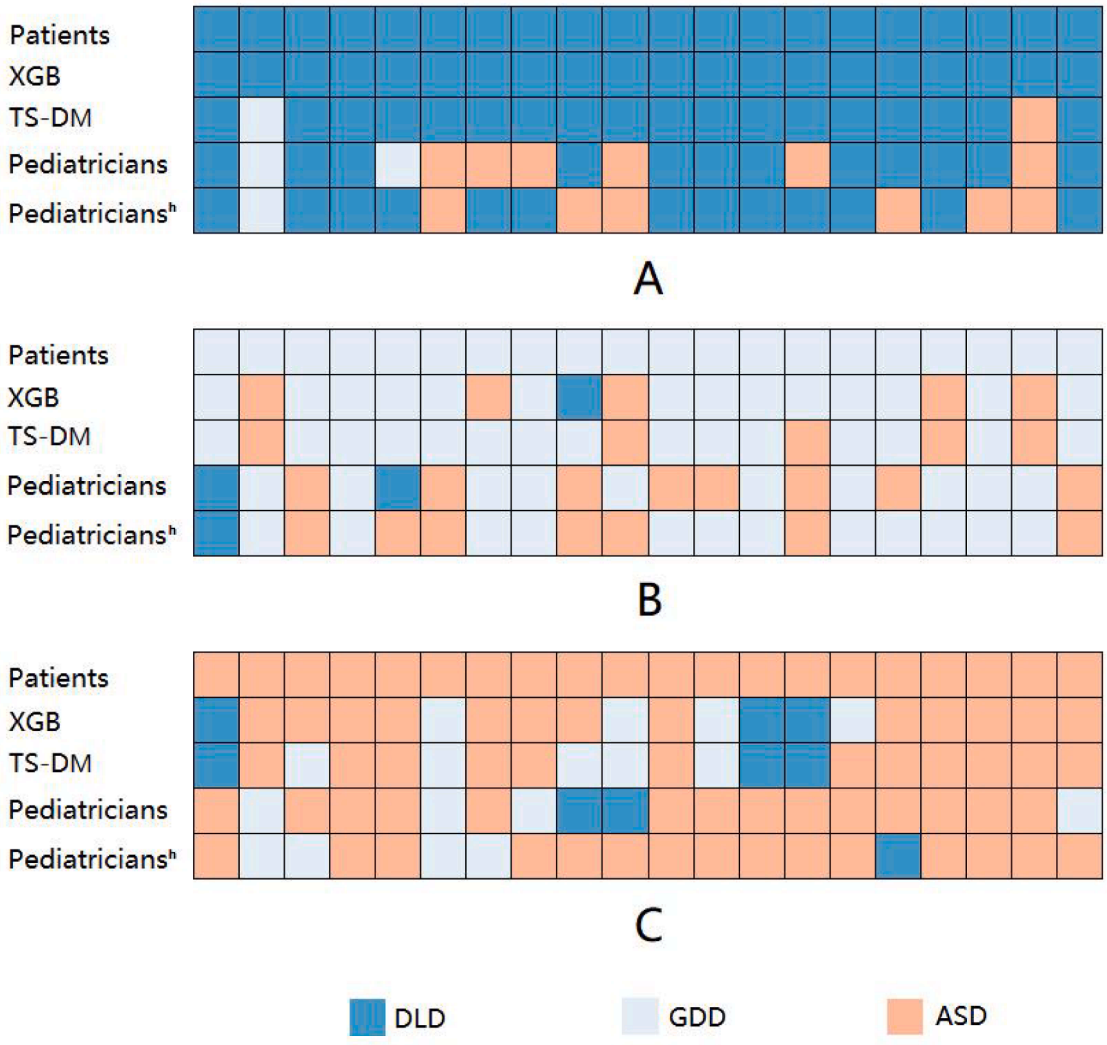

Fig. 3. Results of pediatricians-machine competition on the test dataset of 60 patients.
(A) The patients diagnosed with DLD; (B) The patients diagnosed with GDD; (C) The patients diagnosed with ASD. DLD: Developmental Language Disorder; GDD: Global Developmental Delay; ASD: Autism Spectrum Disorder; Patients: labels of patients' diagnosis; XGB: eXtreme Gradient Boosting; TS-DM: Two-Stage Decision Model; Pediatricians: diagnostic decisions of the pediatricians without using additional medical information; Pediatriciansh: diagnostic decisions of the pediatricians using additional information on patients' medical history.

neurodevelopmental disorders.

In addition to assisting pediatricians in improving diagnostic accuracy, the visible classification process of TS-DM has the potential to increase pediatricians' understanding of the role of the different behavioral and developmental features in the diagnosis process, which may assist them in forming a rigorous clinical mindset and thus

contribute towards more accurate diagnoses. Our findings showed that children with DLD performed better than children with ASD and GDD in adaptive behavior, gross motor, and fine motor abilities, which are consistent with the findings of Saar et al. and Simms et al., suggesting a relative strength of nonverbal cognition in DLD (Saar et al., 2022; Simms and Jin, 2015). For ASD and GDD, the results from the second stage of TS-DM indicated a threshold of three for the number of failed critical items in the M-CHAT to classify ASD and GDD. This threshold value is higher than the cut-off value currently used to screen for ASD in clinical practice, but accords with the findings of a previous study (Bernard et al., 2019), indicating that some children with GDD also manifest some autistic symptoms. Therefore, we suggest that clinicians should be more cautious when diagnosing children with ASD and GDD, especially children with low developmental levels who manifest overlapping symptoms.

In addition, the rankings of important features and the visualized processes TS-DM model showed that nonverbal cognition and autism symptom severity played an important role in classifying the three disorders, whereas language ability seemed to have a limited contribution. This might be because children with any of these three disorders manifest language delay in early childhood (Delehanty et al., 2018), and the similarity of language delay increases the difficulty of distinguishing the disorders. In addition, the participants in this study were toddlers, who are at the early language acquisition stage that involves using a word series without grammar (Al-Harbi, 2020). Although recent studies have demonstrated that morphological and syntactic language, such as pragmatics, syntax, and grammar, may help distinguish ASD (Barsotti et al., 2020), these abilities are more well-developed in older children than in toddlers (Tatlılıoğlu and Senchylo-Tatlilioglu, 2021). Thus, our findings suggest that it is difficult to distinguish these three disorders in early childhood based on language ability, and should rely more on nonverbal cognition and autistic symptoms.

This study had several limitations. First, the developed machine learning model considered only four developmental and behavioral assessment instruments commonly used in China, which would limit its application in countries where other instruments are used. The multi-instrument fusion approach proposed in this paper could be extended to build models based on different instruments in future studies. Second, although the developed model could achieve diagnostic accuracy comparable to that of pediatricians, the test set of patients was from a single center and the sample size was small. Future evaluation of model performance in a multicenter study with a larger sample size is needed. Finally, this study made structural simplifications based on a complex model for better interpretability, resulting in a reduction in the classification accuracy. For clinical applications, an interpretable machine learning model is essential for pediatricians to understand the decision rules and the decision process. Building machine learning models with higher accuracy while ensuring interpretability is a direction for future research.

In summary, we built a machine learning model for the early identification of ASD from DLD and GDD by integrating information from multiple instruments with accuracy similar to that of senior pediatricians. The visualized decision process of TS-DM could increase our understanding of the roles of different behavioral and developmental characteristics in differentiating between these three neurodevelopmental disorders. To facilitate the application of TS-DM, we have created a web application that is freely available at https://pediatricstats.com/tsdm.

Funding/Support

No funding.

Code availability

The machine learning hyperparameters are listed in Supplementary Table 3 and the codes are available at https://github.com/Xuxl2020/ASD-XGB.

CRediT authorship contribution statement

Qiuhong Wei: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization, Validation. **Xueli Xu:** Methodology, Software, Formal analysis, Data curation, Writing – original draft, Visualization, Validation. **Ximing Xu:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration. **Qian Cheng:** Conceptualization, Resources, Writing – review & editing, Supervision, Project administration.

Declaration of Competing Interest

The authors report no conflict of interest.

Data availability

The original data are available on request.

Acknowledgments

We thank Prof. Tingyu Li, Prof. Xiao Liu, and Prof. Ying Dai for reviewing the manuscript and providing clinically relevant suggestions. We also thank the children and their caregivers who participated in this study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.psychres.2023.115050.

References

Al-Harbi, S.S., 2020. Language development and acquisition in early childhood. J. Educ. Learn. 14, 69–73. https://doi.org/10.11591/edulearn.v14i1.14209.

Barsotti, J., Mangani, G., Nencioli, R., Pfanner, L., Tancredi, R., Cosenza, A., et al., 2020. Grammatical comprehension in Italian children with autism spectrum disorder. Brain Sci. 10 https://doi.org/10.3390/brainsci10080510.

Bauer, K., Morin, K.L., Renz III, T.E., Zungu, S., 2022. Autism assessment in low-and middle-income countries: feasibility and usability of western tools. Focus Autism. Dev. Dis. 3, 179–188. https://doi.org/10.1177/10883576211073691.

Bernard, P.M., Mazetto, C., Thiébaut, E., Nassif, M.C., Costa, C.D.S.M., Stefani, A.P., et al., 2019. Heterogeneities in cognitive and socio-emotional development in children with autism spectrum disorder and severe intellectual disability as a comorbidity. Front. Psychiatry 10, 508. https://doi.org/10.3389/fpsyt.2019.00508.

Bishop, D.V.M., Snowling, M.J., Thompson, P.A., Greenhalgh, T., Adams, C., Archibald, L., et al., 2017. Phase 2 of CATALISE: a multinational and multidisciplinary Delphi consensus study of problems with language development: terminology. J. Child Psychol. Psychiatry 58, 1068–1080. https://doi.org/10.1111/jcpp.12721.

Cavus, N., Lawan, A.A., Ibrahim, Z., Dahiru, A., Tahir, S., Abdulrazak, U.I., et al., 2021. A systematic literature review on the application of machine-learning models in behavioral assessment of autism spectrum disorder. J. Pers. Med. 11 https://doi.org/10.3390/jpm11040299.

Chen T., Guestrin C. XGBoost: a scalable tree boosting system; 2016. p. 785–94.

Clark, E., Zhou, Z., Du, L., 2019. Autism in China: progress and challenges in addressing the needs of children and families. Int. J. Sch. Educ. Psychol. 7, 135–146. https://doi.org/10.1080/21683603.2019.1570885.

Collaborators, G.M.D., 2022. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: a systematic analysis for the global burden of disease study 2019. Lancet Psychiat. 9, 137–150. https://doi.org/10.1016/S2215-0366(21)00395-3.

Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G., 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 350, g7594. https://doi.org/10.1136/bmj.g7594.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/bf00994018.

Del, R.C., Slevin, M., Molloy, E.J., Quigley, J., Nixon, E., 2021. How to use the Bayley scales of infant and toddler development. Arch. Dis. Child. Educ. Pract. Ed. 106, 108–112. https://doi.org/10.1136/archdischild-2020-319063.

Delehanty, A.D., Stronach, S., Guthrie, W., Slate, E., Wetherby, A.M., 2018. Verbal and nonverbal outcomesof toddlers with and without autism spectrum disorder, language delay, and global developmental delay. Autism Dev. Lang. Impair. 3, 1875245788 https://doi.org/10.1177/2396941518764764.

Deo, R.C., 2015. Machine learning in medicine. Circulation 132, 1920–1930. https://doi.org/10.1161/CIRCULATIONAHA.115.001593.

Duda, M., Kosmicki, J.A., Wall, D.P., 2014. Testing the accuracy of an observation-based classifier for rapid detection of autism risk. Transl. Psychiatry 4, e424. https://doi.org/10.1038/tp.2014.65.

Fan, R., Hua, T., Shen, T., Jiao, Z., Yue, Q., Chen, B., et al., 2021. Identifying patients with major depressive disorder based on tryptophan hydroxylase-2 methylation using machine learning algorithms. Psychiatry Res. 306, 114258 https://doi.org/10.1016/j.psychres.2021.114258.

Farmer C., Golden C., Thurm A. Concurrent validity of the differential ability scales, second edition with the Mullen Scales of Early Learning in young children with and without neurodevelopmental disorders. *Child Neuropsychol.* 2016;22:556–69. DOI: 10.1080/09297049.2015.1020775.

Genovese, A., Butler, M.G., 2020. Clinical assessment, genetics, and treatment approaches in autism spectrum disorder (ASD). Int. J. Mol. Sci. 21 https://doi.org/10.3390/ijms21134726.

Goel, N., Grover, B., Anuj, Gupta D, Khanna, A., Sharma, M, 2020. Modified grasshopper optimization algorithm for detection of autism spectrum disorder. Phys. Commun.-Amst. 41, 101115 https://doi.org/10.1016/j.phycom.2020.101115.

Hinton, G.E., Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. Science 313, 504–507. https://doi.org/10.1126/science.1127647.

Hus, Y., Segal, O., 2021. Challenges surrounding the diagnosis of autism in children. Neuropsychiatr. Dis. Treat. 17, 3509–3529. https://doi.org/10.2147/NDT.S282569.

Jacob, S., Wolff, J.J., Steinbach, M.S., Doyle, C.B., Kumar, V., Elison, J.T., 2019. Neurodevelopmental heterogeneity and computational approaches for understanding autism. Transl. Psychiat. 9 https://doi.org/10.1038/s41398-019-0390-0.

Larson, A.L., 2016. Language screening for infants and toddlers a literature review of four commercially available tools. Commun. Disord. Q. 1, 3–12. https://doi.org/10.1177/1525740115627420.

Liang, H., Tsui, B.Y., Ni, H., Valentim, C., Baxter, S.L., Liu, G., et al., 2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat. Med. 25, 433–438. https://doi.org/10.1038/s41591-018-0335-9.

Liu, X., Faes, L., Kale, A.U., Wagner, S.K., Fu, D.J., Bruynseels, A., et al., 2019. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit. Health 1, e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2.

Liu, X., Wang, X.M., Ge, J.J., Dong, X.Q., 2018. Effects of the portage early education program on Chinese children with global developmental delay. Medicine 97, e12202. https://doi.org/10.1097/MD.0000000000012202 *(Baltimore)*.

Lord, C., Elsabbagh, M., Baird, G., Veenstra-Vanderweele, J., 2018. Autism spectrum disorder. Lancet 392, 508–520. https://doi.org/10.1016/S0140-6736(18)31129-2.

Lyall, K., Croen, L., Daniels, J., Fallin, M.D., Ladd-Acosta, C., Lee, B.K., et al., 2017. The changing epidemiology of autism spectrum disorders. Annu. Rev. Public Health 38, 81–102. https://doi.org/10.1146/annurev-publhealth-031816-044318.

Maenner, M.J., Shaw, K.A., Bakian, A.V., Bilder, D.A., Durkin, M.S., Esler, A., et al., 2021. Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, United States, 2018. MMWR Surveill. Summ. 70, 1–16. https://doi.org/10.15585/mmwr.ss7011a1.

Marlow, M., Servili, C., Tomlinson, M., 2019. A review of screening tools for the identification of autism spectrum disorders and developmental delay in infants and young children: recommendations for use in low- and middle-income countries. Autism. Res. 12, 176–199. https://doi.org/10.1002/aur.2033.

Miller, L.E., Burke, J.D., Robins, D.L., Fein, D.A., 2019. Diagnosing autism spectrum disorder in children with low mental age. J. Autism. Dev. Disord. 49, 1080–1095. https://doi.org/10.1007/s10803-018-3810-8.

Mithyantha, R., Kneen, R., McCann, E., Gladstone, M., 2017. Current evidence-based recommendations on investigating children with global developmental delay. Arch. Dis. Child. 102, 1071–1076. https://doi.org/10.1136/archdischild-2016-311271.

Mujeeb, R.K., Monica, S.M., 2021. A deep neural network-based model for screening autism spectrum disorder using the quantitative checklist for autism in toddlers (QCHAT). J. Autism Dev. Disord. https://doi.org/10.1007/s10803-021-05141-2.

Ngiam, K.Y., Khor, I.W., 2019. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 20, e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4.

Saar, V., Komulainen, E., Levänen, S., 2022. The significance of nonverbal performance in children with developmental language disorder. Child Neuropsychol. 1–22. https://doi.org/10.1080/09297049.2022.2077324.

Shahamiri, S.R., Thabtah, F., Abdelhamid, N., 2022. A new classification system for autism based on machine learning of artificial intelligence. Technol. Health Care 30, 605–622. https://doi.org/10.3233/THC-213032.

Simms, M.D., Jin, X.M., 2015. Autism, language disorder, and social (Pragmatic) communication disorder: DSM-V and differential diagnoses. Pediatr. Rev. 36, 355–362. https://doi.org/10.1542/pir.36-8-355, 363.

Song, Y.Y., Lu, Y., 2015. Decision tree methods: applications for classification and prediction. Shanghai Arch. Psychiatry 27, 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044.

Steyerberg, E.W., 2019. Clinical Prediction Models. Springer, p. 311, 2nd Version.

Tatlılıoğlu K., Senchylo-Tatlilioglu N. Language development at early childhood: an overview in the context of psycholinguistics. *Psycholinguist. Mod. World* 2021;16: 283–8. DOI: 10.31470/10.31470/2706-7904-2021-16-283-288.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Toh, T.H., Tan, V.W., Lau, P.S., Kiyu, A., 2018. Accuracy of modified checklist for autism in toddlers (M-CHAT) in detecting autism and other developmental disorders in community clinics. J. Autism Dev. Disord. 48, 28–35. https://doi.org/10.1007/s10803-017-3287-x.

Wall, D.P., Kosmicki, J., Deluca, T.F., Harstad, E., Fusaro, V.A., 2012. Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl. Psychiatry 2, e100. https://doi.org/10.1038/tp.2012.10.

Yan, L., Zhang, H.T., Goncalves, J., Xiao, Y., Yuan, Y., 2020. An interpretable mortality prediction model for COVID-19 patients. Nat. Mach. Intell. 2, 283–288. https://doi.org/10.1038/s42256-020-0180-7.

Zhang, C., Chen, X., Wang, S., Hu, J., Wang, C., Liu, X., 2021. Using CatBoost algorithm to identify middle-aged and elderly depression, national health and nutrition examination survey 2011-2018. Psychiatry Res. 306, 114261 https://doi.org/10.1016/j.psychres.2021.114261.

Zhou, H., Xu, X., Yan, W., Zou, X., Wu, L., Luo, X., et al., 2020. Prevalence of autism spectrum disorder in China: a nationwide multi-center population-based study among children aged 6 to 12 Years. Neurosci. Bull. 36, 961–971. https://doi.org/10.1007/s12264-020-00530-6.