



Performance of an artificial intelligence model for classifying stool consistency in infants

Sebastián Palma Mercado¹ , Michelle M. Higuera Carrillo¹ , Johana C. Hincapie Butto²,
Natali González Rozo³ , Carlos Timossi Baldi⁴ , Martha L. Báez Camacho¹ ,
Nicolás I. Ramos Rodríguez¹

ABSTRACT

Introduction. Classifying stool consistency in infants is essential in pediatric gastroenterology. The Brussels Infant and Toddler Stool Scale (BITSS) is the visual standard, but it exhibits interobserver variability. Artificial intelligence (AI) offers an objective alternative to reduce subjectivity.

Objective. To evaluate the agreement between an AI model and the classification performed by pediatric gastroenterologists using BITSS.

Population and methods. An observational, cross-sectional, retrospective study based on a photo database of stool samples from infants under 12 months of age in diapers. We selected 1100 images that met quality criteria; a majority of three gastroenterologists agreed on 896, which we used as the reference set. The images were organized into seven BITSS categories and a grouped version. We trained 20 AI models using an 85/15 split for training and validation. We evaluated performance using exactness, sensitivity, specificity, precision, and F1 score. Inter-specialist agreement was assessed using Fleiss's Kappa, and AI-reference agreement was assessed using weighted Kappa.

Results. The specialists showed 24.7% complete agreement and 75.3% partial agreement, with a Fleiss Kappa of 0.24. The best model for seven categories achieved an overall accuracy of 0.56, with Kappa values of 0.59 (linear) and 0.75 (quadratic). In the grouped version, accuracy reached 0.76, with coefficients of 0.69 and 0.78, respectively.

Conclusion. The AI showed moderate to substantial agreement with the reference classification, particularly in the grouped categories. The findings support the approach's feasibility and warrant external validation studies.

Keywords: artificial intelligence; deep learning; gastroenterology; pediatrics; stool.

doi: <http://dx.doi.org/10.5546/aap.2026-11012.eng>

To cite: Palma Mercado S, Higuera Carrillo MM, Hincapie Butto JC, González Rozo N, Timossi Baldi C, Báez Camacho ML, et al. Performance of an artificial intelligence model for classifying stool consistency in infants. *Arch Argent Pediatr.* 2026;e202611012. Online ahead of print 1-OCT-2026..

¹ Universidad El Bosque, Bogotá, Colombia; ² Hospital Universitario Del Valle, Cali, Colombia; ³ Universidad de Pamplona, Pamplona, Colombia; ⁴ Miramar MedCom LATAM, Mexico.

Correspondence to Sebastián Palma Mercado: spalmam@unbosque.edu.co

Funding: None.

Conflict of interest: None.

Received: 1-4-2026

Accepted: 8-9-2026



This is an open access article under the Creative Commons Attribution–Noncommercial–Noderivatives license 4.0 International. Attribution - Allows reusers to copy and distribute the material in any medium or format so long as attribution is given to the creator. Noncommercial - Only noncommercial uses of the work are permitted. Noderivatives - No derivatives or adaptations of the work are permitted.

INTRODUCTION

Gastrointestinal disorders are a common reason for pediatric visits. It is estimated that 1.7 trillion episodes of diarrhea occur annually in children and that functional constipation has a prevalence ranging from 0.7% to 29.6% in the pediatric population.^{1,2} The pediatric Rome IV criteria for functional constipation classify the condition by age group, distinguishing between toilet-trained and non-toilet-trained patients.³ The assessment of stool consistency is a central component of the diagnostic approach; however, the Bristol Stool Scale, while useful in adults, is not suitable for evaluating stool in diapers.³ For infants, specific scales have been developed, such as the Amsterdam Infant Stool Scale (AISS), validated for children up to 120 days of age, and the Brussels Infant and Toddler Stool Scale (BITSS), created to standardize the assessment of stools in children who do not use the toilet.^{4,5} These tools aim to objectively assess stool consistency, which is relevant in the evaluation of gastrointestinal disorders.

Despite these scales, significant limitations remain in assessing stool in diapers.⁴ Although the use of photographs has improved this process, interobserver variability persists even with standardized scales, leading to diagnostic uncertainty.^{3,5,6}

Recently, various studies have explored artificial intelligence (AI) models based on image recognition to classify stool consistency in children who wear diapers, and these studies showed varying degrees of agreement between photographic assessment, clinical assessment, and caregivers' interpretations.⁶⁻¹⁰

In this context, we evaluated an AI model's performance in classifying stool consistency in infants according to the BITSS scale. We compared it with pediatric gastroenterologists' classifications using concordance analysis in an internal validation setting.

POPULATION AND METHODS

Design and data sources

We conducted an observational, cross-sectional, retrospective agreement study to evaluate agreement between an AI model and pediatric gastroenterologists' consensus in classifying stool consistency in diapers from children under 12 months of age, using the BITSS scale as the reference tool.

The study analyzed photographs of stool samples from diapers of children under 12 months

of age, obtained from a clinical database previously established by the participating team of gastroenterologists. The photographs were taken from preexisting records, with no patient intervention and no individual identifiers.

Inclusion criteria

Feces in diapers of children under 12 months of age. Minimum image resolution: 640 × 480 pixels. Adequate lighting conditions: we included only images with uniform lighting, defined as an average intensity level (R, G, and B) between 80 and 200 on a scale of 0 to 255, based on histogram analysis.

Exclusion criteria

Stool samples collected in a toilet or on a surface other than a diaper. Presence of meconium. Blurred images, images with artifacts, or images of insufficient quality to assess consistency. Lack of consensus: we excluded images if at least two of the three gastroenterologists did not agree on the BITSS classification. Insufficient visualization: images in which less than 50% of the stool sample was visible were excluded. Alteration due to prior handling: we excluded images showing evidence of manual contact, visible alteration, or obvious displacement of the stool sample.

Image quality control

We verified the minimum resolution, lighting conditions, focus, and visual integrity of each image. The average intensity level was analyzed using ImageJ™.

Reference classification

After data cleaning, three pediatric gastroenterologists evaluated the eligible images independently and in a blinded manner, classifying each photograph according to the BITSS scale. The majority vote determined the reference category for each image.

Model training and validation

We trained the model using the Teachable Machine platform, which uses convolutional neural networks, with 85% of the images for training and 15% for internal validation. To explore different configurations, we generated 20 independent models by varying the tool's available training parameters. *Table 1* presents the parameters evaluated for each model.

Statistical analysis

We evaluated the model using overall accuracy and category-specific metrics—sensitivity, specificity, precision, and F1 score—derived from the confusion matrices of the seven- and four-category models.

We analyzed interobserver agreement among the three gastroenterologists using Fleiss’s Kappa and weighted Kappa for pairwise comparisons. We evaluated agreement between the AI model’s classification and that of the specialists using linear and quadratic weighted Kappa, following Landis and Koch’s interpretation criteria. These coefficients were calculated in R™.

Ethical considerations

The Technical Committee of El Bosque University reviewed the protocol on February 6, 2025. In its assessment, the committee determined that the project complies with ethical standards, as it is based on pre-existing, completely anonymous images and involves no interaction with patients or handling of personal information. The study followed the Declaration of Helsinki.

RESULTS

Image selection

The initial set included 1312 photographs, which we reviewed individually against the inclusion and exclusion criteria; of these, we excluded 212 (83 duplicates and 129 that did not meet technical criteria). This left 1100 images eligible for specialist evaluation. We then discarded 204 images due to lack of agreement among evaluators, leaving 896 images with majority agreement among at least two pediatric gastroenterologists. *Figure 1* summarizes the image selection process.

We classified the 896 images according to the BITSS, assigned by consensus among specialists, and distributed them across the scale’s seven original categories (1 through 7). We also established a four-category grouping: constipated (BITSS 1-3), formed (BITSS 4), soft (BITSS 5-6), and liquid (BITSS 7).

Artificial intelligence training

For training and validation, we automatically divided the images according to the platform’s internal schema, preserving the actual distribution of the BITSS categories. We trained 20 independent models (10 for the seven-category scale and 10 for the grouped version), adjusting the available parameters to identify the best configuration. In most models, training accuracy reached 100% between 20 and 35 epochs, with no significant improvement beyond 50, while test-set accuracy remained stable; therefore, we set training to 50 epochs. We then adjusted the learning rate and batch size. *Table 1* presents the accuracy values obtained on the test set; Model 7 (<https://teachablemachine.withgoogle.com/models/9o7XFmnZm/>) showed the best performance in seven-category classification, and Model 19 (<https://teachablemachine.withgoogle.com/models/fwv2rLj6N/>) in clustered classification.

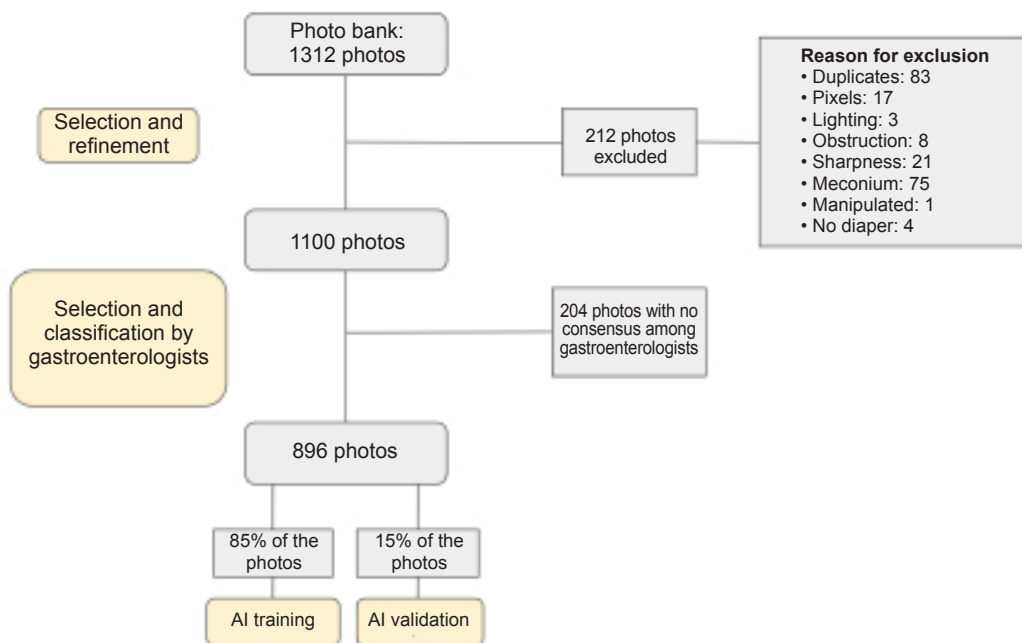
Agreement among evaluators

Among the images classified by the three specialists, evaluators showed varying degrees of agreement. Evaluator 1 and evaluator 2 agreed on 35.7% of the images; evaluator 1 and evaluator 3, on 43.0%; and evaluator 2 and evaluator 3, on 42.9%.

In total, 896 images showed majority agreement. Of these, 675 (75.3%) showed

TABLE 1. Training parameters and accuracy of artificial intelligence models for classification using the Brussels Infant and Toddler Stool Scale (BITSS)

AI Model 7 categories	Size of the batch	Learning rate	Accuracy	AI Model 4 categories	Size of the batch	Learning rate	Accuracy
Model 1	16	0.001	0.47	Model 11	16	0.001	0.69
Model 2	32	0.001	0.50	Model 12	32	0.001	0.67
Model 3	64	0.001	0.48	Model 13	64	0.001	0.67
Model 4	128	0.001	0.56	Model 14	128	0.001	0.69
Model 5	256	0.001	0.53	Model 15	256	0.001	0.68
Model 6	16	0.0005	0.53	Model 16	16	0.0005	0.71
Model 7	32	0.0005	0.566	Model 17	32	0.0005	0.69
Model 8	64	0.0005	0.529	Model 18	64	0.0005	0.71
Model 9	128	0.0005	0.53	Model 19	128	0.0005	0.76
Model 10	256	0.0005	0.54	Model 20	256	0.0005	0.70

FIGURE 1. Photo selection process

AI: artificial intelligence.

partial agreement, and 221 (24.7%) showed total agreement. BITSS 1 had the highest total agreement (48.4%), while BITSS 3 had the lowest (9.3%).

Interobserver agreement (gastroenterologists)

The overall agreement among the three gastroenterologists, assessed using Fleiss's Kappa coefficient, was 0.24. In the pairwise analysis, the linearly and quadratically weighted Kappa coefficients were 0.43 (95% CI: 0.38-0.47) and 0.63 (95% CI: 0.59-0.68) between evaluators 1 and 2; 0.55 (95% CI: 0.52-0.58) and 0.76 (95% CI: 0.73-0.78) between raters 1 and 3; and 0.51 (95% CI: 0.48-0.55) and 0.70 (95% CI: 0.65-0.73) between raters 2 and 3.

Performance of the artificial intelligence model

For classification according to the seven-category BITSS scale, Model 7 performed best, with an overall accuracy of 0.56. Sensitivity ranged from 0.11 (BITSS 4) to 0.86 (BITSS 5). Specificity reached 1.00 in BITSS categories 1, 2, and 4, while in the others it ranged from 0.55 (BITSS 5) to 0.98 (BITSS 3). Precision ranged from 0.50 to 1.00, and the F1 score

ranged from 0.20 (BITSS 4) to 0.89 (BITSS 1).

In the four-category classification, Model 19 achieved an overall accuracy of 0.76. Sensitivity was highest in the hard-stool group (0.93), followed by the soft-stool group (0.86) and the liquid-stool group (0.65), with the lowest value in the formed-stool group (0.39). Specificity remained high in most groups (0.95–0.98), except in the soft-stool group (0.67). Precision ranged from 0.54 (formed stools) to 0.87 (constipated stools), and the F1 score ranged from 0.45 to 0.90, with the best performance also observed in the constipated stool group.

Table 2 presents the sensitivity, specificity, precision, and F1 score metrics for models 7 and 19.

Alignment between artificial intelligence and gastroenterologists

We assessed agreement between the AI model and pediatric gastroenterologists using weighted Kappa (linear and quadratic) based on the confusion matrices of the seven- and four-category BITSS models shown in Figure 2.

For the seven-category model, the linear-weighted Kappa was 0.59 (95% CI: 0.47-0.69), and the quadratic-weighted Kappa was 0.75 (95% CI: 0.61-0.84). In the four-category grouping, the

values were higher, with a linear-weighted Kappa of 0.69 (95% CI: 0.55-0.78) and a quadratic-weighted Kappa of 0.78 (95% CI: 0.68-0.87).

DISCUSSION

This study evaluated an AI model’s performance in classifying stool consistency in infants using the BITSS scale and assessed agreement among pediatric gastroenterologists as part of an internal validation study. This study provides evidence from a Latin American context and highlights both the limitations of visual assessment and AI’s potential to reduce diagnostic variability.

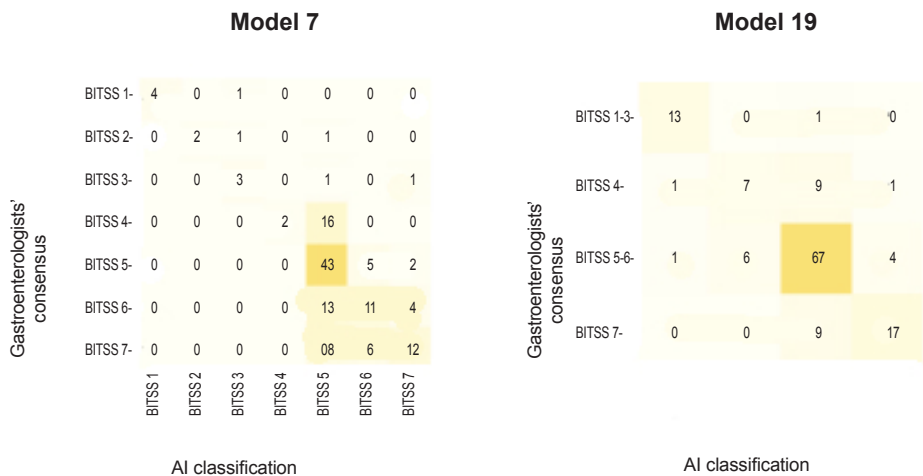
Gastroenterologists’ evaluations showed significant variability in classification according to the BITSS scale, even though there was majority agreement on all images. Unanimity among the three evaluators was rare, occurring in about one-quarter of the images, confirming the difficulty of distinguishing certain categories, particularly BITSS 3 and 4, which have less distinct visual features and greater interpretive discrepancy. This observation was reflected in the concordance analysis, where most discrepancies occurred between adjacent categories—a pattern previously described by Aman et al. in fresh specimens and photographic evaluations.⁶

TABLE 2. Performance metrics by category for models 7 and 19

BITSS category (Model 7)	Sensitivity (recall)	Specificity	Precision	F1 score	BITSS category (Model 19)	Sensitivity (recall)	Specificity	Precision	F1 score
BITSS 1	0.80	1	1	0.89	Constipation (BITSS 1-3)	0.93	0.98	0.87	0.90
BITSS 2	0.50	1	1	0.67					
BITSS 3	0.60	0.98	0.60	0.60					
BITSS 4	0.11	1	1	0.20	Formed (BITSS 4)	0.39	0.95	0.54	0.45
BITSS 5	0.86	0.55	0.52	0.65	Soft (BITSS 5-6)	0.86	0.67	0.78	0.82
BITSS 6	0.39	0.90	0.50	0.44					
BITSS 7	0.46	0.94	0.63	0.53	Liquid (BITSS 7)	0.65	0.95	0.77	0.70

BITSS: Brussels Infant and Toddler Stool Scale.

FIGURE 2. Confusion matrices (Models 7 and 19)



BITSS: Brussels Infant and Toddler Stool Scale; AI: artificial intelligence.

In this context, the low Fleiss Kappa value (0.24) reflects low agreement among the three raters. At the same time, the weighted Kappa coefficients were higher because they accounted for the scale's ordinal nature and penalized disagreements between adjacent categories to a lesser extent.

In the seven-category AI model, performance followed a pattern similar to that of the specialists. The BITSS 1 category (constipated) was one of the most easily recognized by the algorithm, as was BITSS 5, which corresponded to the class with the most images in the training set. In contrast, BITSS 4 remained the category with the lowest discriminatory power, suggesting that its visual features are less distinct and reflect the ambiguity present in clinical evaluation. The concentration of predictions in adjacent categories indicates that the model identified shared characteristics among closely related classes, influenced by morphological similarity and uneven distribution. This pattern aligns with studies on ordinal classification in medical images, where errors tend to cluster in adjacent categories.¹¹

Literature supports these findings. Ludwig et al. reported an overall accuracy of nearly 60% in a model trained on a larger dataset, suggesting that classification difficulties depend not solely on database size but on the intrinsic complexity of the full BITSS scale.⁷

The four-category scale performed better than the seven-category version. Constipated stools were identified most consistently, followed by soft stools, which maintained adequate overall performance despite lower specificity. In contrast, formed stools (BITSS 4) remained the worst-performing category, mirroring the pattern seen in the seven-category model. This persistent difficulty is consistent with previous reports by other authors, in which the functional grouping of BITSS categories was associated with greater accuracy and less variability.^{5,7}

The agreement between the AI and the reference classification was moderate to substantial, with higher values on the grouped scale, suggesting that reducing the number of categories facilitates the algorithm's discrimination and improves its stability. This aligns with the scale's ordinal nature, where transitions between levels are gradual and create ambiguity for both human raters and automated models. Previous studies have shown that the differences between adjacent categories are minimal, with extensive

overlap and greater difficulty at the intermediate levels.^{12,13}

Although most of the model's errors were concentrated in adjacent categories, broader misclassifications were also observed, possibly related to the imbalance in the number of images per category and the presence of intermediate features within the same image—a phenomenon previously described even among trained observers.¹⁴ Since neural network-based models require large, balanced datasets, future versions should increase the representation of less frequent categories.^{15,16} However, visual evaluation of deposits has inherent limitations and subjectivity, and uneven class distribution can skew performance toward the most represented categories, as previously described and observed in this study.¹⁷

The Teachable Machine platform limits access to technical model details, which restricts full reproducibility of the training process. Furthermore, the image database did not allow us to establish independence by patient or the individual contribution of each infant. Excluding images without consensus and applying minimum photographic quality criteria may have reduced the dataset's representativeness compared with typical clinical scenarios. Future studies could evaluate the model's performance on subsets with unanimous agreement to explore its stability under different levels of diagnostic certainty. Taken together, these results underscore the complexity of fecal classification in infants and the need for external validation before clinical application.

CONCLUSION

The BITSS scale shows inter-observer variability, particularly in intermediate categories, and the artificial intelligence model performed comparably to specialists, with greater stability when using the grouped version. These results provide evidence of internal validation and preliminary reproducibility; therefore, they should not be interpreted as proof of clinical transferability. The findings support the feasibility of this approach and justify its evaluation in external validation studies. ■

REFERENCES

1. World Health Organization. Diarrhoeal disease. Geneva: WHO; 2024. [Accessed on November 16, 2025]. Available from: <https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease>
2. Koppen IJN, Velasco-Benítez CA, Benninga MA, Di Lorenzo

- C, Saps M. Using the Bristol Stool Scale and parental report of stool consistency as part of the Rome III criteria for functional constipation in infants and toddlers. *J Pediatr*. 2016;177:44-8.e1. doi:10.1016/j.jpeds.2016.06.055.
3. Velasco-Benítez CA, Llanos-Chea A, Saps M. Utility of the Brussels Infant and Toddler Stool Scale (BITSS) and Bristol Stool Scale in non-toilet-trained children: a large comparative study. *Neurogastroenterol Motil*. 2021;33(8):e14015. doi:10.1111/nmo.14015.
 4. Silva LCD, Bianchini PM, Ortolan EVP, Hamamoto JF, Fermiano R, Rego RMP, et al. Brazilian Portuguese version of the Amsterdam Infant Stool Scale: a valid and reliable scale for evaluation of stool from children up to 120 days old. *BMC Pediatr*. 2021;21(1):64. doi:10.1186/s12887-021-02527-0.
 5. Huysentruyt K, Koppen I, Benninga M, Cattaert T, Cheng J, De Geyter C, et al. The Brussels Infant and Toddler Stool Scale: A study on interobserver reliability. *J Pediatr Gastroenterol Nutr*. 2019;68(2):207-13. doi:10.1097/MPG.0000000000002153.
 6. Aman BA, Levy EI, Hofman B, Vandenplas Y, Huysentruyt K. Real time versus photographic assessment of stool consistency using the Brussels Infant and Toddler Stool Scale: Are they telling us the same? *Pediatr Gastroenterol Hepatol Nutr*. 2021;24(1):38-44. doi:10.5223/pghn.2021.24.1.38.
 7. Ludwig T, Oukid I, Wong J, Ting S, Huysentruyt K, Roy P, et al. Machine learning supports automated digital image scoring of stool consistency in diapers. *J Pediatr Gastroenterol Nutr*. 2021;72(2):255-61. doi:10.1097/MPG.0000000000003007.
 8. Wu J, Dong L, Sun Y, Zhao X, Gan J, Wang Z. The effectiveness of artificial intelligence in assisting mothers with assessing infant stool consistency in a breastfeeding cohort study in China. *Nutrients*. 2024;16(6):855. doi:10.3390/nu16060855.
 9. Pimentel M, Mathur R, Wang J, Chang C, Hosseini A, Fiorentino A, et al. A smartphone application using artificial intelligence is superior to subject self-reporting when assessing stool form. *Am J Gastroenterol*. 2022;117(7):1118-24. doi:10.14309/ajg.0000000000001723.
 10. Xiao F, Wang Y, Ludwig T, Li X, Chen S, Sun N, et al. Generation and application of a convolutional neural networks algorithm in evaluating stool consistency in diapers. *Acta Paediatr*. 2023;112(6):1333-40. doi:10.1111/apa.16731.
 11. Liu X, Zou Y, Song Y, Yang C, You J, Kumar BVKV. Ordinal regression with neuron stick-breaking for medical diagnosis. In: Leal-Taixé L, Roth S (Eds). *Computer Vision – ECCV 2018 Workshops, Proceedings*. Springer Verlag; 2019:335-44. doi:10.1007/978-3-030-11024-6_23.
 12. Blake MR, Raker JM, Whelan K. Validity and reliability of the Bristol Stool Form Scale in healthy adults and patients with diarrhoea-predominant irritable bowel syndrome. *Aliment Pharmacol Ther*. 2016;44(7):693-703. doi:10.1111/apt.13746.
 13. Zhou J, McNabb J, DeCapite N, Ruiz JR, Fisher DA, Grego S, et al. Stool image analysis for digital health monitoring by smart toilets. *IEEE Internet Things J*. 2023;10(5):3720-34. doi:10.1109/JIOT.2022.3158886.
 14. Bekkali N, Hamers SL, Reitsma JB, Van Toledo L, Benninga MA. Infant Stool Form Scale: development and results. *J Pediatr*. 2009;154(4):521-6.e1. doi:10.1016/j.jpeds.2008.10.010.
 15. Zhang Z, Zhang X, Ichiji K, Bukovský I, Homma N. How intra-source imbalanced datasets impact the performance of deep learning for COVID-19 diagnosis using chest X-ray images. *Sci Rep*. 2023;13(1):19049. doi:10.1038/s41598-023-45368-w.
 16. Salmi M, Atif D, Oliva D, Abraham A, Ventura S. Handling imbalanced medical datasets: review of a decade of research. *Artif Intell Rev*. 2024;57:273. doi:10.1007/s10462-024-10884-2.
 17. Choy YP, Hu G, Chen J. Detection and classification of human stool using deep convolutional neural networks. *IEEE Access*. 2021;9:160485-96. doi:10.1109/ACCESS.2021.3131982.