



Desempeño de un modelo de inteligencia artificial para la clasificación de la consistencia de las deposiciones en lactantes

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 

RESUMEN

Introducción. La clasificación de la consistencia de las deposiciones en lactantes es fundamental en gastroenterología pediátrica. La Brussels Infant and Toddler Stool Scale (BITSS) es el estándar visual, pero presenta variabilidad interobservador. La inteligencia artificial (IA) ofrece una alternativa objetiva para reducir la subjetividad.

Objetivo. Evaluar la concordancia entre un modelo de IA y la clasificación realizada por gastroenterólogos pediatras utilizando BITSS.

Población y métodos. Estudio observacional, transversal y retrospectivo basado en un banco fotográfico de deposiciones en pañales de lactantes menores de 12 meses. Se seleccionaron 1100 imágenes que cumplieran criterios de calidad; 896 alcanzaron acuerdo por mayoría entre tres gastroenterólogos y se emplearon como referencia. Las imágenes se organizaron en siete categorías de BITSS y una versión agrupada. Se entrenaron veinte modelos de IA mediante división 85/15 para entrenamiento y validación. Se evaluó el desempeño mediante exactitud, sensibilidad, especificidad, precisión y puntuación F1. La concordancia entre especialistas se evaluó con Kappa de Fleiss y la concordancia IA-referencia, con Kappa ponderado.

Resultados. Los especialistas presentaron un acuerdo total del 24,7 % y parcial del 75,3 %, con un Kappa de Fleiss de 0,24. El mejor modelo para siete categorías obtuvo una exactitud global de 0,56, con Kappa de 0,59 (lineal) y 0,75 (cuadrático). En la versión agrupada, la exactitud alcanzó 0,76, con coeficientes de 0,69 y 0,78, respectivamente.

Conclusión. La IA mostró concordancia moderada a sustancial con la clasificación de referencia, especialmente en las categorías agrupadas. Los hallazgos respaldan la factibilidad del enfoque y justifican estudios de validación externa.

Palabras clave: inteligencia artificial; aprendizaje profundo; gastroenterología; pediatría; heces.

doi (español): <http://dx.doi.org/10.5546/aap.2026-11012>

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

Cómo citar: Palma Mercado S, Higuera Carrillo MM, Hincapie Butto JC, González Rozo N, Timossi Baldi C, Báez Camacho ML, et al. Desempeño de un modelo de inteligencia artificial para la clasificación de la consistencia de las deposiciones en lactantes. *Arch Argent Pediatr.* 2026;e202611012. Primero en Internet 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, México.

Correspondencia para Sebastián Palma Mercado: spalmam@unbosque.edu.co

Financiamiento: Ninguno.

Conflicto de intereses: Ninguno que declarar.

Recibido: 4-1-2026

Aceptado: 9-8-2026



Esta obra está bajo una licencia de Creative Commons Atribución-No Comercial-Sin Obra Derivada 4.0 Internacional. Atribución — Permite copiar, distribuir y comunicar públicamente la obra. A cambio se debe reconocer y citar al autor original. No Comercial — Esta obra no puede ser utilizada con finalidades comerciales, a menos que se obtenga el permiso. Sin Obra Derivada — Si remezcla, transforma o crea a partir del material, no puede difundir el material modificado.

INTRODUCCIÓN

Los trastornos gastrointestinales representan un motivo de consulta frecuente en pediatría. Se estima que ocurren 1,7 billones de episodios de diarrea anuales en niños y que el estreñimiento funcional presenta una prevalencia entre el 0,7 % y el 29,6 % en la población infantil.^{1,2} Los criterios Roma IV pediátricos de estreñimiento funcional dividen la clasificación según grupos etarios, diferenciando entre pacientes entrenados y no entrenados en el uso de inodoro.³ La evaluación de la consistencia de las heces es un componente central en la aproximación diagnóstica; sin embargo, la escala de Bristol, aunque útil en adultos, no es adecuada para evaluar heces en pañales.³ Para los lactantes, se desarrollaron escalas específicas, como la Amsterdam Infant Stool Scale (AISS), validada para niños hasta 120 días de vida, y la Brussels Infant and Toddler Stool Scale (BITSS), creadas para estandarizar la evaluación de deposiciones en niños que no utilizan el inodoro.^{4,5} Estas herramientas buscan objetivar la consistencia fecal, relevante en la evaluación de trastornos gastrointestinales.

Pese a las escalas existentes, persisten limitaciones importantes en la evaluación de las deposiciones en pañales.⁴ Aunque el uso de fotografías ha mejorado este proceso, subsiste variabilidad interobservador incluso con escalas estandarizadas, lo que genera incertidumbre diagnóstica.^{3,5,6}

Recientemente, distintos estudios han explorado modelos de inteligencia artificial (IA) basados en reconocimiento de imágenes para clasificar la consistencia fecal en niños que utilizan pañales, y mostraron acuerdos variables entre la evaluación fotográfica, la evaluación clínica y la interpretación de cuidadores.⁶⁻¹⁰

En este contexto, se evaluó el desempeño de un modelo de IA para clasificar la consistencia fecal en lactantes según la escala BITSS, en comparación con la clasificación de gastroenterólogos pediatras, mediante análisis de concordancia en un escenario de validación interna.

POBLACIÓN Y MÉTODOS

Diseño y fuente de datos

Se realizó un estudio observacional, transversal y retrospectivo de concordancia, orientado a evaluar el grado de acuerdo entre un modelo de IA y el criterio conjunto de gastroenterólogos pediatras en la clasificación de la consistencia de las heces en pañales de

niños menores de 12 meses, utilizando la escala BITSS como herramienta de referencia.

El estudio analizó imágenes fotográficas de deposiciones en pañales de niños menores de 12 meses, obtenidas de un banco de datos clínico previamente establecido por el equipo de gastroenterólogos participantes. Las fotografías correspondían a registros preexistentes, sin intervención sobre los pacientes ni identificadores individuales.

Criterios de inclusión

Deposiciones en pañal de niños menores de 12 meses. Resolución mínima de la imagen 640 × 480 píxeles. Condiciones adecuadas de iluminación: se incluyeron únicamente imágenes con iluminación homogénea, definida como nivel promedio de intensidad (R, G y B) entre 80 y 200 en escala de 0 a 255, según análisis del histograma.

Criterios de exclusión

Deposiciones registradas en inodoro u otra superficie distinta a un pañal. Presencia de meconio. Imágenes desenfocadas, con artefactos o con calidad insuficiente para evaluar la consistencia. Ausencia de consenso: se excluyeron las imágenes en las que al menos dos de los tres gastroenterólogos no alcanzaron acuerdo en la clasificación BITSS. Visualización insuficiente: se excluyeron las imágenes en las que se apreciaba menos del 50 % de la deposición. Alteración por manipulación previa: se excluyeron imágenes con evidencia de contacto manual, alteración visible o desplazamiento evidente de la deposición.

Control de calidad de las imágenes

Se verificó la resolución mínima, las condiciones de iluminación, el enfoque y la integridad visual de cada imagen. El nivel promedio de intensidad se analizó mediante ImageJ.

Clasificación de referencia

Tras la depuración, las imágenes elegibles fueron evaluadas por tres gastroenterólogos pediatras, de manera independiente y ciega, quienes clasificaron cada fotografía según la escala BITSS. La categoría de referencia para cada imagen se definió por mayoría.

Entrenamiento y validación del modelo

El modelo se entrenó utilizando la plataforma Teachable Machine, basada en redes neuronales

convolucionales, empleando un 85 % de las imágenes para el entrenamiento y un 15 % para la validación interna. Para explorar distintas configuraciones, se generaron 20 modelos independientes mediante la variación de los parámetros de entrenamiento disponibles en la herramienta. Los parámetros evaluados para cada modelo se presentan en la *Tabla 1*.

Análisis estadístico

El modelo se evaluó mediante exactitud global y métricas por categoría: sensibilidad, especificidad, precisión y puntuación F1, obtenidas a partir de las matrices de confusión de los modelos de siete y cuatro categorías.

La concordancia interobservador entre los tres gastroenterólogos se analizó mediante el coeficiente Kappa de Fleiss y los coeficientes Kappa ponderados para las comparaciones por pares. La concordancia entre la clasificación del modelo de IA y los especialistas se evaluó utilizando Kappa ponderado lineal y cuadrático, siguiendo los criterios de interpretación de Landis y Koch. El cálculo de estos coeficientes se realizó en el *software R*.

Consideraciones éticas

El protocolo fue revisado por el Comité Técnico de la Universidad El Bosque el 6 de febrero de 2025. En su concepto, el proyecto cumple con los aspectos éticos, dado que se basa en imágenes preexistentes, completamente anónimas, sin interacción con pacientes ni manejo de información personal. El estudio se desarrolló conforme a los principios de la Declaración de Helsinki.

RESULTADOS

Selección de imágenes

El conjunto inicial incluyó 1312 fotografías, revisadas individualmente según criterios de inclusión y exclusión, de las cuales se excluyeron 212 (83 duplicadas y 129 por no cumplir criterios técnicos). Quedaron 1100 imágenes elegibles para la evaluación por especialistas. Posteriormente, se descartaron 204 imágenes por falta de acuerdo entre evaluadores; se obtuvo un conjunto final de 896 imágenes con acuerdo mayoritario entre al menos dos gastroenterólogos pediatras. La *Figura 1* resume el proceso de selección de imágenes.

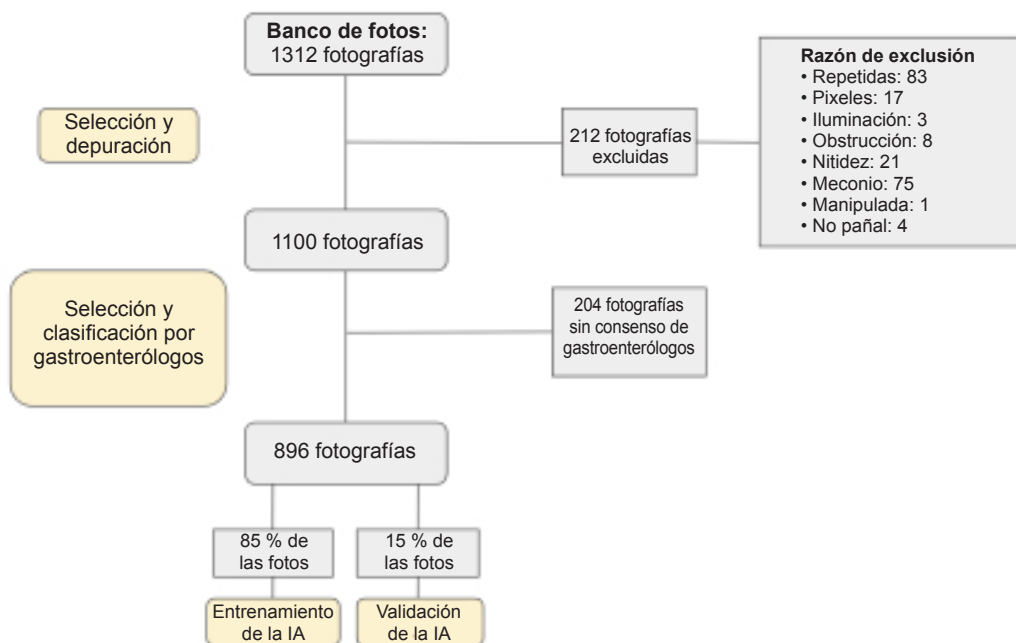
Las 896 imágenes fueron clasificadas según BITSS, asignada por consenso entre los especialistas, distribuyéndose en las siete categorías originales de la escala (1 a 7). Además, se estableció una agrupación adicional en cuatro categorías: Estreñimiento (BITSS 1-3), Formadas (BITSS 4), Blandas (BITSS 5-6) y Líquidas (BITSS 7).

Entrenamiento de la inteligencia artificial

Para el entrenamiento y validación, las imágenes se dividieron automáticamente según el esquema interno de la plataforma, conservando la distribución real de las categorías BITSS. Se entrenaron 20 modelos independientes (10 para la escala de siete categorías y 10 para la versión agrupada), ajustando los parámetros disponibles para identificar la mejor configuración. En la mayoría de los modelos, la exactitud de entrenamiento alcanzó el 100 % entre 20 y 35 épocas, sin mejoras relevantes más allá de las 50, mientras que la exactitud del conjunto

TABLA 1. Parámetros de entrenamiento y exactitud de los modelos de inteligencia artificial para la clasificación Brussels Infant and Toddler Stool Scale (BITSS)

Modelo IA 7 categorías	Tamaño de lote	Tasa de aprendizaje	Exactitud (Accuracy)	Modelo IA 4 categorías	Tamaño de lote	Tasa de aprendizaje	Exactitud (Accuracy)
Modelo 1	16	0,001	0,47	Modelo 11	16	0,001	0,69
Modelo 2	32	0,001	0,50	Modelo 12	32	0,001	0,67
Modelo 3	64	0,001	0,48	Modelo 13	64	0,001	0,67
Modelo 4	128	0,001	0,56	Modelo 14	128	0,001	0,69
Modelo 5	256	0,001	0,53	Modelo 15	256	0,001	0,68
Modelo 6	16	0,0005	0,53	Modelo 16	16	0,0005	0,71
Modelo 7	32	0,0005	0,566	Modelo 17	32	0,0005	0,69
Modelo 8	64	0,0005	0,529	Modelo 18	64	0,0005	0,71
Modelo 9	128	0,0005	0,53	Modelo 19	128	0,0005	0,76
Modelo 10	256	0,0005	0,54	Modelo 20	256	0,0005	0,70

FIGURA 1. Proceso de selección de fotografías

de prueba se mantuvo estable; por ello, se fijó el entrenamiento en 50 épocas. Los ajustes posteriores se centraron en la tasa de aprendizaje y el tamaño del lote. La *Tabla 1* presenta los valores de exactitud obtenidos en el conjunto de prueba; el Modelo 7 (<https://teachablemachine.withgoogle.com/models/9o7XFmnZm/>) mostró el mejor desempeño en la clasificación de siete categorías y el Modelo 19 (<https://teachablemachine.withgoogle.com/models/fwv2rLj6N/>) en la clasificación agrupada.

Acuerdo entre evaluadores

De las imágenes clasificadas por los tres especialistas, se observó un grado variable de coincidencia entre pares de evaluadores. Evaluador 1 y evaluador 2 coincidieron en el 35,7 % de las imágenes; evaluador 1 y evaluador 3, en el 43,0 %; y evaluador 2 y evaluador 3, en el 42,9 %.

En total, 896 imágenes presentaron acuerdo por mayoría. De estas, 675 (75,3 %) correspondieron a acuerdo parcial y 221 (24,7 %) a acuerdo total. La categoría con mayor acuerdo total fue BITSS 1 (48,4 %), mientras que la categoría con menor acuerdo total fue BITSS 3 (9,3 %).

Concordancia interobservador (gastroenterólogos)

La concordancia global entre los tres gastroenterólogos, evaluada mediante el coeficiente Kappa de Fleiss, fue de 0,24. En el análisis por pares, los coeficientes Kappa ponderados lineal y cuadrático fueron de 0,43 (IC95% 0,38-0,47) y 0,63 (IC95% 0,59-0,68) entre los evaluadores 1 y 2; de 0,55 (IC95% 0,52-0,58) y 0,76 (IC95% 0,73-0,78) entre los evaluadores 1 y 3; y de 0,51 (IC95% 0,48-0,55) y 0,70 (IC95% 0,65-0,73) entre los evaluadores 2 y 3.

Desempeño del modelo de inteligencia artificial

Para la clasificación según la escala BITSS de siete categorías, el modelo con mejor desempeño fue el Modelo 7, con una exactitud global de 0,56. La sensibilidad varió entre 0,11 (BITSS 4) y 0,86 (BITSS 5). La especificidad alcanzó 1,00 en las categorías BITSS 1, 2 y 4, mientras que en las demás osciló entre 0,55 (BITSS 5) y 0,98 (BITSS 3). La precisión se ubicó entre 0,50 y 1,00; y la puntuación F1, entre 0,20 (BITSS 4) y 0,89 (BITSS 1).

En la clasificación agrupada en cuatro categorías, el Modelo 19 obtuvo una exactitud global de 0,76. La sensibilidad fue mayor en el

grupo de heces estreñidas (0,93), seguida de las blandas (0,86) y las líquidas (0,65), con el valor más bajo en las formadas (0,39). La especificidad se mantuvo elevada en la mayoría de los grupos (0,95-0,98), excepto en el grupo de heces blandas (0,67). La precisión varió entre 0,54 (formadas) y 0,87 (estreñidas), y la puntuación F1 entre 0,45 y 0,90, con su mejor desempeño también en el grupo estreñido.

La *Tabla 2* presenta las métricas de sensibilidad, especificidad, precisión y puntuación F1 para los modelos 7 y 19.

Concordancia entre la inteligencia artificial y los gastroenterólogos

La concordancia entre el modelo de IA y los gastroenterólogos pediatras se evaluó mediante Kappa ponderado (lineal y cuadrático) a partir de las matrices de confusión de los modelos BITSS de siete y cuatro categorías presentadas en la *Figura 2*.

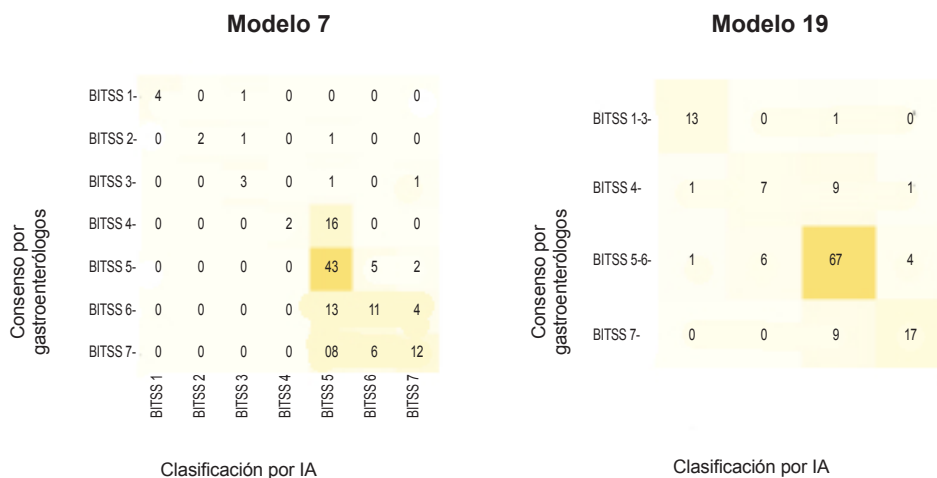
Para el modelo de siete categorías, el Kappa ponderado lineal fue de 0,59 (IC95% 0,47-0,69) y el Kappa ponderado cuadrático de 0,75 (IC95% 0,61-0,84). En la clasificación agrupada en cuatro categorías, los valores fueron superiores, con un Kappa ponderado lineal de 0,69 (IC95% 0,55-

TABLA 2. Métricas de desempeño por categoría para los Modelos 7 y 19

Categoría BITSS (Modelo 7)	Sensibilidad (recall)	Especificidad	Precisión	Puntuación F1	Categoría BITSS (Modelo 19)	Sensibilidad (recall)	Especificidad	Precisión	Puntuación F1
BITSS 1	0,80	1	1	0,89	Estreñimiento (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	Formadas (BITSS 4)	0,39	0,95	0,54	0,45
BITSS 5	0,86	0,55	0,52	0,65	Blandas (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	Líquidas (BITSS 7)	0,65	0,95	0,77	0,70

BITSS: Brussels Infant and Toddler Stool Scale.

FIGURA 2. Matrices de confusión (Modelo 7 y 19)



BITSS: Brussels Infant and Toddler Stool Scale; IA: inteligencia artificial.

0,78) y un Kappa ponderado cuadrático de 0,78 (IC95% 0,68-0,87).

DISCUSIÓN

Este estudio evaluó el desempeño de un modelo de IA para clasificar la consistencia fecal en lactantes mediante la escala BITSS y la concordancia entre gastroenterólogos pediatras, en el marco de una validación interna. Este trabajo contribuye con evidencia en un contexto latinoamericano y resalta tanto las limitaciones de la evaluación visual como el potencial de la IA para reducir la variabilidad diagnóstica.

La evaluación de los gastroenterólogos mostró una variabilidad relevante en la clasificación según la escala BITSS, aun cuando todas las imágenes contaban con acuerdo mayoritario. La unanimidad entre los tres evaluadores fue infrecuente y se concentró en aproximadamente una cuarta parte de las imágenes, lo que confirma la dificultad para discriminar ciertas categorías, particularmente BITSS 3 y 4, caracterizadas por rasgos visuales menos definidos y mayor discrepancia interpretativa. Esta observación se reflejó en la concordancia, donde la mayoría de los desaciertos ocurrió entre categorías adyacentes, un patrón descrito previamente por Aman *et al.* en muestras frescas y evaluaciones fotográficas.⁶

En este contexto, el bajo valor del Kappa de Fleiss (0,24) se explica por la baja concordancia entre los tres evaluadores, mientras que los coeficientes Kappa ponderados fueron más altos al considerar la naturaleza ordinal de la escala y penalizar en menor medida los desacuerdos entre categorías adyacentes.

En el modelo de IA de siete categorías, el rendimiento mostró un patrón similar al de los especialistas. La categoría BITSS 1 (estreñida) fue una de las más fácilmente reconocidas por el algoritmo, al igual que BITSS 5, que correspondía a la clase con mayor número de imágenes en el conjunto de entrenamiento. En contraste, BITSS 4 se mantuvo como la categoría con menor capacidad de discriminación, lo que sugiere que sus rasgos visuales son menos definidos y reproducen la ambigüedad presente en la evaluación clínica. La concentración de predicciones en categorías adyacentes indica que el modelo identificó características compartidas entre clases próximas, influido por la similitud morfológica y la distribución desigual. Este patrón concuerda con estudios de clasificación ordinal en imágenes médicas, donde los errores suelen

concentrarse en categorías adyacentes.¹¹

La literatura respalda estos hallazgos. Ludwig *et al.* reportaron una exactitud global cercana al 60 % en un modelo entrenado con un conjunto de imágenes mayor, lo que sugiere que las dificultades de clasificación no dependen únicamente del tamaño de la base de datos, sino de la complejidad intrínseca de la escala BITSS completa.⁷

La escala agrupada en cuatro categorías mostró mejor desempeño frente a la versión de siete categorías. Las heces estreñidas fueron las más consistentemente identificadas, seguidas por las blandas, que mantuvieron un rendimiento global adecuado pese a menor especificidad. En contraste, las heces formadas (BITSS 4) continuaron siendo la categoría con peor desempeño, reproduciendo el patrón del modelo de siete categorías. Esta dificultad persistente coincide con lo reportado previamente por otros autores, donde la agrupación funcional de categorías BITSS se asoció con mayor exactitud y menor variabilidad.^{5,7}

La concordancia entre la IA y la clasificación de referencia fue de moderada a sustancial, con valores superiores en la escala agrupada, lo que sugiere que reducir el número de categorías facilita la discriminación del algoritmo y mejora su estabilidad. Esto es coherente con la naturaleza ordinal de la escala, en la que las transiciones entre niveles son graduales y generan ambigüedad tanto en evaluadores humanos como en modelos automatizados. Estudios previos han demostrado que las diferencias entre categorías adyacentes son mínimas, con amplia superposición y mayor dificultad en los niveles intermedios.^{12,13}

Aunque la mayoría de los errores del modelo se concentraron en categorías adyacentes, también se observaron desaciertos más amplios, posiblemente relacionados con el desequilibrio en el número de imágenes por categoría y con la presencia de características intermedias en una misma deposición, un fenómeno descrito previamente incluso entre observadores entrenados.¹⁴ Dado que los modelos basados en redes neuronales requieren conjuntos amplios y balanceados, futuras versiones deberían aumentar la representación de las categorías menos frecuentes.^{15,16} No obstante, la evaluación visual de deposiciones presenta limitaciones inherentes y subjetividad, y la distribución desigual de clases puede sesgar el rendimiento hacia las categorías más representadas, como

se ha descrito previamente y se observó en este estudio.¹⁷

El uso de la plataforma Teachable Machine limita el acceso a detalles técnicos del modelo, lo que restringe la reproducibilidad completa del proceso de entrenamiento. Asimismo, la naturaleza del banco de imágenes no permitió establecer la independencia por paciente ni la contribución individual de cada lactante. La exclusión de imágenes sin consenso y los criterios mínimos de calidad fotográfica pueden haber reducido la representatividad del conjunto de datos frente a escenarios clínicos habituales. Futuros estudios podrían evaluar el desempeño del modelo en subconjuntos con acuerdo unánime para explorar su estabilidad bajo distintos niveles de certeza diagnóstica. En conjunto, estos resultados refuerzan la complejidad de la clasificación fecal en lactantes y la necesidad de validación externa antes de su aplicación clínica.

CONCLUSIÓN

La escala BITSS muestra variabilidad interobservador, especialmente en categorías intermedias, y el modelo de inteligencia artificial alcanzó un desempeño comparable al de los especialistas, con mayor estabilidad al utilizar la versión agrupada. Estos resultados corresponden a evidencia de validación interna y reproducibilidad preliminar del modelo, por lo que no deben interpretarse como demostración de transportabilidad clínica. Los hallazgos apoyan la factibilidad de este enfoque y justifican su evaluación en estudios de validación externa. ■

REFERENCIAS

- World Health Organization. Diarrhoeal disease. Geneva: WHO; 2024. [Consulta: 16 de noviembre de 2025]. Disponible en: <https://www.who.int/news-room/fact-sheets/detail/diarrhoeal-disease>
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.