Calidad y particionamiento de fotografía clínica no estandarizada en aprendizaje supervisado

José Pablo Soriano-Torres[1 [0009−0008−3150−2528]], Isabel María López-Medina[2 [0000−0002−3437−9229]], Francisco Charte[1 [0000−0002−3083−8942]], César Hueso-Montoro[2 [0000−0003−1515−3870]], Carmen Álvarez-Nieto[2 [0000−0002−0913−7893]], Concepción Capilla-Díaz[3 [0000−0002−3361−8718]], and María Dolores Pérez-Godoy[1 [0000−0002−6670−564X]]

[1] Departamento de Informática, Universidad de Jaén, Jaén, España
{psoriano,fcharte,lperez}@ujaen.es
[2] Departamento de Enfermería, Universidad de Jaén, Jaén, España
{imlopez,chueso,calvarez}@ujaen.es
[3] Departamento de Enfermería, Universidad de Granada, Granada, España
{conchicd}@ugr.es

Resumen La adopción de la fotografía clínica mediante dispositivos móviles (*Smartphone-Based Imaging*) ofrece grandes oportunidades para la monitorización telemática en dermatología u ostomías, entre otras, pero carece de la estandarización propia de los entornos radiológicos tradicionales. Esta variabilidad intrínseca induce sesgos instrumentales, morfológicos e institucionales que propician el aprendizaje de atajos (*shortcut learning*). Sin embargo, la literatura actual suele abordar estas imágenes utilizando técnicas de particionamiento de datos que provocan una severa sobreestimación del rendimiento clínico. En este artículo proponemos un marco metodológico para aislar y medir la verdadera capacidad de generalización de los modelos de *deep learning*. Nuestra propuesta combina arquitecturas de particionado avanzado, como el aislamiento estricto de pacientes y la validación multicéntrica (*Leave-One-Center-Out*), junto con auditorías de Inteligencia Artificial Explicable (Grad-CAM). Validando empíricamente el marco sobre un caso de estudio de complicaciones en piel periestomal, demostramos cómo el particionamiento estándar sufre de fuga de datos (*data leakage*), incrementando artificialmente el F1-Score a un 85.1 %, frente a la realidad clínica del 52.1 % obtenida bajo aislamiento real. Los resultados evidencian que el diseño de la validación es un requisito predictivo del éxito clínico, superando en relevancia a la complejidad de la red.

Keywords: Deep Learning · Particionamiento de datos · Fotografía clínica móvil · Fuga de datos (*Data Leakage*) · Aprendizaje de atajos (*Shortcut Learning*) · Inteligencia Artificial Explicable (XAI).

1. Introducción

El diagnóstico por imagen es un pilar fundamental para el análisis clínico y la intervención médica. Tradicionalmente, esta disciplina ha operado bajo entornos altamente controlados, apoyándose en estándares robustos como *Digital Imaging and Communication in Medicine* (DICOM) [15]. Este tipo de estándares garantizan la uniformidad, interoperabilidad y consistencia en modalidades de captura de imágenes médicas estructuradas.

Sin embargo, la generalización de los dispositivos móviles ha impulsado la adopción de la fotografía clínica en consulta, conocida como ***smartphone-based imaging* (SBI)** [10]. Esta práctica es especialmente valiosa en dermatología, seguimiento de heridas, análisis de la piel periestomal, entre otras, pero obliga a los modelos de *machine learning* (ML) y *deep learning* (DL) a trabajar con imágenes que no siguen ningún estándar predefinido. Esta ausencia de normalización introduce una serie de desafíos críticos provocando problemas en la generalización de modelos de inteligencia artificial (IA).

A pesar de la urgencia por integrar estas herramientas en la práctica clínica, la literatura reciente ha priorizado el diseño de nuevas arquitecturas de red, frente a la gestión y particionamiento de los datos [19]. Sin embargo, esta presunción da lugar a varios problemas: obvia por completo las interdependencias generadas por la morfología del paciente, las características propias del hardware de los dispositivos y las condiciones ambientales propias de cada institución médica. En consecuencia, el estado del arte actual sufre frecuentemente de una sobreestimación del rendimiento, en la que los modelos aparentan ser precisos al haber memorizado atajos (*shortcut learning* [5]) en lugar de aprender características inherentes a los datos.

Para dar respuesta a esta problemática, este artículo tiene como propósito **definir y validar** empíricamente un enfoque **metodológico**, orientado al tratamiento de imágenes médicas no estandarizadas (SBI), aunque puede extenderse a otro tipo de imágenes médicas.

El resto de este artículo se estructura de la siguiente manera. La Sección 2 establece el marco teórico, revisando el compromiso sesgo-varianza y las limitaciones de las estrategias de aumentación clásicas en entornos clínicos. La Sección 3 expone el estado del arte respecto a la taxonomía de la variabilidad en fotografía no controlada. La Sección 4 delimita nuestra contribución principal, presentando el marco metodológico de particionamiento e interpretabilidad visual. La Sección 5 materializa empíricamente nuestra propuesta mediante dos casos de estudio sobre diagnóstico en piel periestomal. Finalmente, la Sección 6 destila estos aprendizajes en una guía de buenas prácticas, culminando con las conclusiones en la Sección 7.

2. Antecedentes

La migración desde imágenes DICOM hacia capturas móviles introduce un cambio en los paradigmas tradicionales del aprendizaje automático. Para fundamentar metodológicamente las estrategias que propondremos, resulta importante

comprender cómo reaccionan las redes neuronales ante distribuciones ruidosas y desbalanceadas.

2.1. Evaluación: el compromiso sesgo-varianza en entornos ruidosos

En la práctica clínica en abierto, los protocolos de estandarización fotográfica rara vez se cumplen de manera estricta. Para evitar que las redes neuronales colapsen memorizando el ruido contextual en lugar de la patología subyacente, el diseño del conjunto de datos cobra un papel protagonista frente a la arquitectura. En ML, el pilar sobre el que se sustenta la capacidad de predicción y generalización es el rigor en el particionamiento de la información [16]. Clásicamente, este se divide en:

- **Training set**. Usado para calcular y actualizar los pesos internos del modelo.
- **Validation set**. Usado para llevar a cabo un seguimiento del entrenamiento, ajustar hiperparámetros y determinar criterios de parada (*early stopping*).
- **Test set**. Su uso es exclusivo para la inferencia y evaluación final de la capacidad de generalización del modelo, debiendo permanecer inalterado e invisible hasta dicho momento.

El particionamiento modula directamente uno de los axiomas más importantes del modelado predictivo: el **compromiso entre sesgo y varianza** [3]. El sesgo representa la incapacidad del algoritmo para capturar la verdadera relación entre las características y el diagnóstico, mientras que la varianza mide la hipersensibilidad del modelo ante cambios en el conjunto de entrenamiento.

2.2. Desbalanceo en datos médicos y las fronteras de la aumentación

Habitualmente, los datos médicos presentan categorías con muchos más datos que otras. El enfoque tradicional para mitigar este desbalance han sido las estrategias de *data augmentation* basadas en transformaciones espaciales (rotaciones, traslaciones) y de color. No obstante, en la visión médica estas alteraciones pueden provocar cambios no válidos.

Por ello, se han desarrollado diversos enfoques que operan sobre la función de pérdida y la síntesis semántica, elementos que deben complementar a un diseño particional limpio:

- **Aprendizaje sensible al coste (*Cost-Sensitive Learning*)**: penaliza asimétricamente los errores, inyectando un multiplicador de impacto en la *Weighted Cross-Entropy* para forzar a la retropropagación (*backpropagation*) a proteger las detecciones minoritarias críticas [8].
- ***Focal Loss***: popularizada en entornos de alta densidad, esta modulación de la entropía cruzada atenúa la contribución de los ejemplos bien clasificados de la clase mayoritaria sana, focalizando la actualización de pesos en ejemplos difíciles o atípicos de la frontera de decisión [13].
- **Generación avanzada (GANs y modelos de difusión)**: el estado del arte ha migrado hacia la inyección de información sintética patológicamente coherente [12]. Sin embargo, validar el impacto real de esta síntesis requiere, de nuevo, un entorno de test puramente independiente e inalterado.

3. Taxonomía de la variabilidad en entornos no controlados

Aplicar modelos supervisados sobre imágenes médicas no estandarizadas resulta un reto complejo debido a la gran variabilidad de las mismas. Como resume la Tabla 1, el conocimiento existente categoriza esta variabilidad en tres ejes fundamentales: dispositivo, institución y paciente.

Tabla 1: Taxonomía de la variabilidad en fotografía clínica y su impacto documentado en el ciclo de vida del modelo.

Origen	**Desafío Técnico**	**Efecto en el Modelo**	**Estrategia de Mitigación Previa**
Dispositivo	Distorsión de lente, procesado por software, fondo.	*Shortcut learning* (sesgo por instrumentación).	Control de captura [11], máscaras.
Institución	Hardware dispar, PACS heterogéneos, protocolos.	Caída drástica de rendimiento (*domain shift*).	Estandarización de iluminación y color [22].
Paciente	Oclusiones (vello, cicatrices), sesgo demográfico.	Sesgo discriminatorio, falsos positivos por anatomía.	Aumentación y balanceo [7].

3.1. Variabilidad debida al dispositivo

Una barrera crítica en SBI radica en las diferencias de hardware entre fabricantes. Distintas lentes producen distorsiones geométricas variables, y el procesado de imagen interno (*ISP pipeline*) provoca que la textura de un tejido luzca diferente según el móvil que la capture [11]. Adicionalmente, el factor humano sin protocolos de estandarización [20] altera la escala y perspectiva espaciales. Esto fomenta el aprendizaje de atajos (*shortcut learning*), asociando falsamente marcadores espurios (una regla milimetrada médica, marcas de rotulador quirúrgico) al diagnóstico clínico [21,5].

3.2. Variabilidad debido a la institución

El cambio de dominio institucional (*inter-institutional domain shift*) [24,23] invalida las soluciones locales. El entorno lumínico (ej. lámparas LED cálidas frente a luz blanca de quirófano) y la infraestructura digital (algoritmos de compresión específicos de los PACS hospitalarios [6]) inyectan una firma característica en las imágenes. Si estas variables correlacionan por azar con la prevalencia en el diagnóstico, el algoritmo asociará el origen hospitalario a la gravedad del cuadro, induciendo graves fallos entre hospitales.

3.3. Variabilidad debido al paciente

La biología individual dificulta el trazado de fronteras de decisión puras. Más allá del sesgo demográfico histórico en la literatura (como el rendimiento inferior sobre fototipos oscuros [7]), las peculiaridades topológicas –vello, cicatrices

previas, variaciones en la vascularización– modifican abruptamente los mapas de activación [2]. Un estoma con protrusión prominente en estado sano puede confundirse fácilmente con una inflamación anómala si dicha fisonomía no está suficientemente representada en la distribución de entrenamiento, causando alta tasa de falsos positivos.

4. Marco metodológico de particionamiento y validación visual

Frente al estado del arte que trata las variaciones clínicas como un problema puramente de aumentación o robustez del algoritmo, la **contribución principal de este trabajo** consiste en articular y formalizar una solución preventiva en la capa de datos. Para aumentar la fiabilidad diagnóstica, proponemos el siguiente marco metodológico estructurado que fusiona estrategias topológicas de particionado con auditoría XAI. Nuestro objetivo no es presentar un modelo algorítmico, sino una metodología que mitigue la sobreestimación en el rendimiento de modelos IA sobre imágenes no estandarizadas.

4.1. Estrategias de particionado topológico clínico

Para garantizar que las métricas reflejen generalización real, es necesario sustituir los enfoques de partición aleatoria por arquitecturas que respeten el aislamiento causal de las variables. Planteamos esta progresión estructural:

- ***Hold-out* estratificado (*Stratified Split*)**: dado el severo desbalanceo, garantiza que la prevalencia patológica sea la misma en *train*, *validation* y *test*, estabilizando el cálculo del VPP (Valor predictivo positivo) [1].
- **Aislamiento iterativo de pacientes (*Stratified Group Shuffle Split*)**: en SBI y monitorización continua (como la de las ostomías), conviven decenas de capturas del mismo paciente. Separarlas corrompe el experimento, pues la red no aprende la enfermedad, sino que memoriza la morfología identificativa del individuo compartida entre *train* y *test*. El aislamiento requiere emplear el ID del paciente como ancla invariable para la separación [17], reteniendo a su vez las proporciones de prevalencia.
- **Validación multicéntrica (*Leave One Center Out - LOCO*)**: supone una prueba interesante frente al *domain shift* institucional (Sección 3.2). El modelo se expone iterativamente a $N-1$ centros hospitalarios y se evalúa exclusivamente contra un centro aislado [23]. Es uno de los diseños capaz de certificar si las representaciones del modelo poseen valor médico extrapolable a otro centro hospitalario.

4.2. Auditoría y purificación mediante IA explicable (XAI) y modelos fundacionales

La segregación de los datos requiere una validación cruzada cualitativa. Las arquitecturas profundas son opacas; por ello, la interpretabilidad post-hoc actúa como una garantía empírica en la fase de prueba.

Mediante la extracción de mapas de activación con **Grad-CAM** [18], los médicos pueden certificar visualmente si las zonas específicas usadas por los modelos de IA convergen exactamente sobre la lesión biológica. Si la red concentra su atención en márgenes de una gasa, se invalida el modelo a pesar de sus altas métricas.

Para mantener coherente el marco predictivo frente al entorno y el dispositivo, nuestra metodología prescribe la segmentación anatómica previa. Aprovechar modelos fundacionales como el *Segment Anything Model* (SAM) o su refinamiento clínico **MedSAM** [14] permite extraer la región de interés (ROI), eliminando el 90 % del ruido externo antes de alimentar a las redes clasificadoras.

5. Validación empírica: impacto del particionamiento en la evaluación sobre piel periestomal

Para demostrar empíricamente que con la metodología propuesta se evita la sobreestimación de los modelos, se ha diseñado una validación sobre un conjunto de datos propietario de imágenes de ostomías y complicaciones de la piel periestomal. A continuación, describiremos el dataset.

El dataset está formado por 1 980 imágenes. Estas están divididas en cuatro categorías según la escala DET [4]:

- Escala DET 0. En esta categoría hay 187 imágenes.
- Escala DET 1-3. En esta categoría hay 617 imágenes.
- Escala DET 4-7. En esta categoría hay 743 imágenes.
- Escala DET 8-15. En esta categoría hay 433 imágenes.

De todo el conjunto de imágenes, podemos obtener diferente información acerca de **los pacientes** y de **los hospitales** introducidos en este estudio.

En cuanto a los **pacientes**. Hay un total de 306 pacientes únicos. Esto quiere decir que en el conjunto de datos hay más de una imagen por paciente; de manera resumida, la media de imágenes por paciente es de 6.47 imágenes con una desviación típica de 2.10 imágenes. De la primera categoría, de 187 imágenes, hay 35 pacientes únicos. De la segunda, de 617 imágenes, hay 87 pacientes únicos. De la tercera, de 743 imágenes, hay 113 pacientes únicos. De la última, de 433 imágenes, hay 70 pacientes únicos.

En cuanto a los **hospitales**. Hay un total de 17 hospitales únicos. En la Tabla 2, podemos apreciar la distribución de imágenes por categoría en cada hospital. Se puede apreciar que hospitales como HUSC poseen el 12 % de las imágenes del conjunto de datos, mientras que otros como el HSJC posee únicamente el 0.45 %.

Este conjunto muestra muchos de los desafíos de la fotografía de dispositivo móvil en entorno ambulatorio. Reafirmamos que la intención no es propones un nuevo estado del arte algorítmico, sino poner de manifiesto la necesidad de una metodología de particionamiento que evite la sobreestimación del rendimiento de modelos que realmente no consiguen una capacidad tan alta de generalización.

Tabla 2: porcentaje de imágenes por hospital dada la categoría a la que pertenecen.

Categoría/Hospital	DET 0 (%)	DET 1-3 (%)	DET 4-7 (%)	DET 8-15 (%)	Total (%)
HA	0.66	0.51	1.52	0.51	3.18
HLL	0.45	0.00	0.66	0.66	1.77
HMA	0.66	3.74	1.77	1.11	7.27
HP	0.00	0.05	0.15	0.10	0.30
HSJC	0.25	0.00	0.20	0.00	0.45
HUJ	1.06	5.30	5.20	1.41	12.98
HUJF	0.71	1.26	3.89	2.12	7.98
HUPM	0.00	0.35	1.06	1.36	2.78
HURS	1.11	2.58	5.71	2.53	11.92
HUS	0.25	0.30	0.61	0.3	1.46
HUSA	1.62	6.92	2.12	1.06	11.72
HUSC	1.52	3.79	3.99	2.78	12.07
HUT	0.20	0.91	0.00	0.35	1.46
HUVM	0.35	1.01	3.18	2.42	6.97
HUVV	0.25	0.51	0.30	0.91	1.97
HVN	0.00	1.92	2.98	1.06	5.96
HVR	0.35	2.02	4.19	3.18	9.75
Total	9.44	31.16	37.53	21.87	100.00

5.1. Caso de estudio: importancia del particionamiento en un problema de clasificación multiclase

En este caso de estudio, realizaremos un entrenamiento y una prueba sobre el conjunto de datos descrito anteriormente. Primero mostraremos el valor de las diferentes métricas tras un particionamiento que no tiene en cuenta el paciente ni hospital y, después, los resultados cuando sí se tiene en cuenta la separación.

En primer lugar, partimos de una red convolucional DenseNet121[9] preentrenada con el conjunto de datos ImageNet. Se ha fijado una semilla con valor 42. Para la aumentación de datos, hemos incrementado en un 20 % la clase mayoritaria y balanceado el resto de clases hasta igualar dicho volumen de imágenes, aplicando principalmente las siguientes técnicas: rotación, recorte, traslación y distorsión.

El entrenamiento se ha configurado con un tamaño de lote de 32 y 50 épocas sobre las tres últimas capas de la arquitectura: el clasificador, el bloque denso 4 y la capa de normalización 5. Se ha empleado un planificador de la tasa de aprendizaje de tipo coseno con 10 épocas de calentamiento, una función de pérdida de entropía cruzada y el optimizador AdamW, con una tasa de aprendizaje de 0.0005 y una decaída de pesos de 0.001. Cabe destacar que, sobre el conjunto de datos utilizado, se ha realizado una segmentación previa para quedarnos únicamente con la piel periestomal, obviando tanto el fondo como el estoma.

Posteriormente, se ha ejecutado un entrenamiento con todas las imágenes excepto las pertenecientes a los hospitales HUJF, HA y HUVV, que se han mantenido como conjunto de prueba.

Por tanto, hemos obtenido dos resultados tras la finalización del entrenamiento:

- La evaluación del modelo sobre un conjunto de validación, obtenido mediante un muestreo sin reemplazo del 10 % del conjunto de entrenamiento, obviando la posible coincidencia de pacientes y hospitales entre ambos conjuntos.
- La evaluación del modelo sobre el conjunto de prueba, donde nos hemos cerciorado de no incluir ningún paciente ni hospital que ya estuviera presente en el conjunto de entrenamiento.

Así, los resultados obtenidos se pueden apreciar en la Tabla 3. El enfoque que mezcla pacientes y hospitales reporta unas métricas muy buenas (F1 del 85.1 %). Sin embargo, este rendimiento es artificial: ocurre un fenómeno grave de *data leakage*. La red deduce atajos fácilmente reconocibles si el individuo pertenece parcialmente a la memoria del *train*. Al implementar la agrupación por paciente y por hospital, forzamos al modelo a funcionar correctamente ante pacientes nunca vistos. La consecuente y notable caída de las métricas (el F1 se desploma al 52.1 %) indica que realmente, el modelo no tiene una capacidad de generalización tan alta como lo que se esperaba.

Tabla 3: Comparativa del rendimiento métrico en el problema de clasificación multiclase.

Estrategia de Particionado	**Precisión (Acc) %**	**F1-Score Macro %**	**AUC %**	**MCC**
Sin separar pacientes y hospitales	86.6	85.1	96.6	0.81
Separando pacientes y hospitales	**53.1**	**52.1**	**77.5**	**0.367**

Para validar esta hipótesis de memorización morfológica, se ha recurrido a la inspección mediante interpretabilidad post-hoc. La Figura 1 presenta los mapas de activación Grad-CAM extraídos de las predicciones sobre una imagen de un paciente perteneciente al conjunto de prueba. El modelo se ha fijado en la zona inferior izquierda de la piel periestomal que presenta un tono diferente al resto para predecir que esta imagen pertenece al grupo DET 8-15, cuando la realidad es todo lo contrario. Se puede observar que una evaluación más robusta evidencia que el modelo tiene dificultades para fijarse en las zonas realmente importantes.

En conclusión, entrenar un modelo con imágenes médicas (concretamente del tipo SBI, aunque se podría extrapolar a otras imágenes clínicas) presenta un gran desafío. Este problema se debe a la obtención de **métricas muy optimistas** al entrenar y evaluar simultáneamente con imágenes que proceden del mismo paciente y hospital. La drástica **caída de rendimiento** posterior demuestra que, al realizar un particionamiento más estricto, complicamos la tarea del modelo, obligándolo a fijarse realmente en las zonas clínicas relevantes en lugar de en patrones de fondo inherentes al paciente o al centro de salud. También es cierto que las métricas descienden a valores modestos, aunque en este punto entran en juego factores como la naturaleza del propio conjunto de datos usado y el

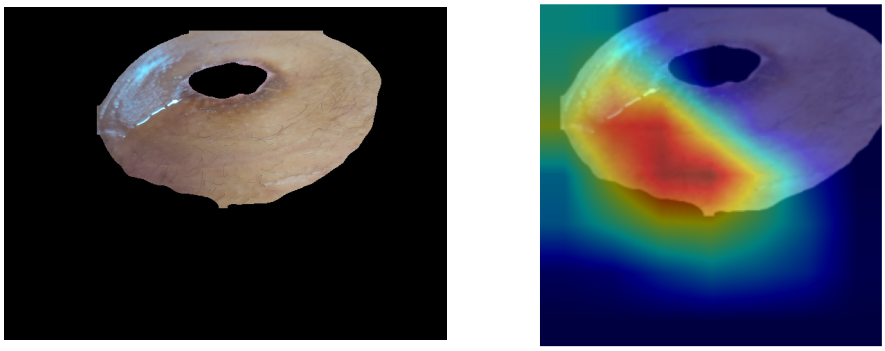


Figura 1: En la izquierda podemos ver la imagen original, mientras que a la derecha se muestra la misma imagen tras aplicar Grad-CAM. Esta imagen pertenece al conjunto de prueba que se aisló considerando pacientes y hospitales. Su categoría real es DET 0, mientras que ha sido predicha como DET 8-15.

desbalanceo de imágenes por clase, características que recaen fuera del ámbito de este estudio.

6. Guía de estandarización desde la adquisición hasta la evaluación

Para facilitar la adopción práctica y asegurar la reproducibilidad metodológica del ecosistema biomédico, presentamos la siguiente rúbrica de cuatro fases. Se recomienda su inclusión en la sección material y métodos de futuras publicaciones basadas en fotografía de *smartphones*:

Fase 1: adquisición y control de calidad

- **Instrumental y entorno:** especificación rigurosa del hardware de captura (ej. modelos de *smartphone*), el uso de luz (LED, natural, xenón) y estabilización espacial.
- **Exclusión visual experta:** documentación de la tasa de descarte de imágenes desenfocadas, o con un contraste tan bajo que un clínico humano no lograría un consenso diagnóstico.
- **Normalización geométrica:** inclusión obligatoria de patrones de referencia (reglas estándar en el borde de la captura) que posibiliten el re-escalado durante el entrenamiento. Así como mantener el tamaño, la orientación y la relación de aspecto idénticas en todas las imágenes.

Fase 2: curación y criba computacional

- **Filtrado de variables de confusión (*confounders*):** aplicación de modelos como MedSAM para asegurar que elementos irrelevantes (grapas, ropa, joyas, sondas) no inyecten una varianza ficticia durante el cálculo del gradiente en el tejido de interés.
- **Armonización en el color:** justificación de la estrategia de corrección lumínica empleada para solventar la intervención del procesado *software* de las cámaras de los móviles sobre los tonos reales de la piel.

Fase 3: elección de particionamiento de datos

- **Aislamiento del paciente:** implementación de *Stratified Group Shuffle Split* para garantizar la contención de todas las imágenes temporales de un paciente en un único conjunto de evaluación.
- **Aislamiento de hospitales:** evaluación en un hospital u hospitales no vistos en el entrenamiento (*Leave-One-Center-Out*).

Fase 4: explicabilidad y depuración de sesgos

- **Correlación anatómica visual:** inserción de análisis cualitativos de la capa de decisión final (ej. mapas térmicos Grad-CAM) confirmando que la atención otorgada por el modelo se relaciona con el motivo de la predicción.

7. Conclusiones

La fotografía móvil (*Smartphone-Based Imaging*) representa un horizonte prometedor para escalar la teledermatología en ostomías y heridas complejas. Sin embargo, nuestra investigación demuestra que la flexibilidad de estas capturas penaliza la capacidad de generalización de los algoritmos de DL.

Por consiguiente, hemos mostrado que el diseño meticuloso de la validación experimental es esencial. Las métricas optimistas obtenidas mediante particiones aleatorias ocultan fugas de información (*data leakage*). La solidez de un estudio depende del aislamiento estricto por paciente, la validación interinstitucional (estrategia LOCO) y la purificación previa de las imágenes mediante segmentación con modelos fundacionales.

8. Agradecimientos

Este estudio recibió apoyo financiero de múltiples agencias de financiación. Por un lado, forma parte del proyecto “Avances en el desarrollo de modelos de IA confiables para contribuir a la adopción y uso de una IA responsable en Salud (TAIH)” , con el código PID2023-149511OB-I00, financiado por MICIU/AEI /10.13039/50110001 1033 y por ERDF, UE. Por otro lado, forma parte del proyecto de investigación “Avances en la evaluación de la piel periestomal en

personas con ostomías de eliminación digestiva. Uso de Modelos de Inteligencia Artificial (EPPIA)". M.1.B.B TA_000115_UJA23 cofinanciada por Consejería de Universidad, Investigación e Innovación y por Programa FEDER Andalucía 2021-2027.

Referencias

1. Altman, D.G., Bland, J.M.: Diagnostic tests 2: Predictive values. BMJ: British Medical Journal **309**(6947), 102 (1994)
2. Bissoto, A., Valle, E., Avila, S.: Deconstructing bias in skin lesion datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 726–727 (2020)
3. Briscoe, E., Feldman, J.: Conceptual complexity and the bias/variance tradeoff. Cognition **118**(1), 2–16 (2011)
4. Coloplast: Escala DET: Herramienta de valoración de la piel periestomal
5. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (Nov 2020)
6. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations (2019)
7. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1820–1828 (2021)
8. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert systems with applications **73**, 220–239 (2017)
9. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks (2018)
10. Hunt, B., Ruiz, A., Pogue, B.: Smartphone-based imaging systems for medical applications: a critical review. Journal of biomedical optics **26**(4), 040902 (2021)
11. Jayasuriya, N., Munasinghe, B.M., Ravihari, K., Sirimanna, S., Jayalath, M., Weerasundara, G.: More than just a click: A five-step guide to mobile medical photography for amateurs. Cureus **17**(8), e89668 (aug 2025)
12. Kazeminia, S., Baur, C., Kuijf, H.J., Menze, B.H., Engelhardt, S., et al.: Gans for medical image analysis. Artificial Intelligence in Medicine **109**, 101938 (2020)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 2980–2988 (2017)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
15. Mildenberger, P., Eichelberg, M., Martin, E.: Introduction to the dicom standard. European radiology **12**(4), 920–927 (2002)
16. Reitermanova, Z., et al.: Data splitting. In: WDS. vol. 10, pp. 31–36. Matfyzpress Prague (2010)
17. Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C., Kording, K.P.: The need to avoid data leakage in machine learning-based translational medicine. GigaScience **6**(5), gix022 (2017)

18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
19. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. npj Digital Medicine **5**(1), 1–8 (2022)
20. Wang, W.H., Liu, G.T., Chen, S.J., Chiang, Y.C., Wang, H.H., Chiang, Y.C., tzu Hu, Y., Hsu, W.S.: Artificial intelligence assisted smartphone system for real-time detection and severity assessment of digital clubbing. International Journal of Medical Sciences **23**, 1613 – 1629 (2026)
21. Winkler, J.K., Fink, C., Toberer, F., Enk, A., Haenssle, H.A.: Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA dermatology **155**(10), 1135–1141 (2019)
22. Witmer, W.K., Lebovitz, P.J.: Clinical photography in the dermatology practice. Seminars in cutaneous medicine and surgery **31**(3), 191–199 (2012)
23. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Medicine **15**(11), e1002683 (2018)
24. Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Wood, B.J., Roth, H., Myronenko, A., Xu, D., et al.: Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. IEEE transactions on medical imaging **39**(7), 2531–2540 (2020)