

NEURODERM: SISTEMA BASADO EN REDES NEURONALES CONVOLUCIONALES PARA LA CLASIFICACION DE SEVERIDAD DE DERMATITIS ATÓPICA

Pablo Israel López Ballinas ^a, Gerardo Ferras Toledo ^a.

^aEstudiantes de Ingeniería Biomédica, Universidad Politécnica de Chiapas.

Autor por correspondencia: 231080@ib.upchiapas.edu.mx

RESUMEN

La dermatitis atópica es una enfermedad cutánea crónica de alta prevalencia en la población infantil y juvenil. Su diagnóstico tradicional, predominantemente visual, presenta limitaciones de subjetividad en etapas tempranas. Este trabajo tiene como objetivo desarrollar un sistema inteligente basado en redes neuronales profundas para apoyar el diagnóstico y la clasificación de severidad de esta afección. El alcance comprende una herramienta de asistencia digital orientada al primer nivel de atención médica en contextos con recursos sanitarios limitados. La metodología consistió en el procesamiento y clasificación multiclase de imágenes dermatológicas mediante transferencia de aprendizaje con la arquitectura DenseNet201, adaptando un enfoque inicialmente binario hacia un clasificador del gradiente clínico. Los resultados indican que el sistema alcanzó una exactitud global (accuracy) del 77.04% y un AUC- ROC macro del 91.34% en el conjunto de prueba. El análisis por categoría evidenció un alto desempeño en los estados extremos de la patología,

registrando una sensibilidad (recall) del 81.82% para la clase leve y del 83.96% para la clase grave. La categoría moderada, aunque presentó mayor complejidad debido a fronteras fenotípicas difusas, alcanzó una sensibilidad del 58.16%.

En conclusión, el modelo demuestra una sólida robustez diagnóstica y representa una alternativa tecnológica eficaz para mitigar la variabilidad en el criterio médico en entornos con escasa infraestructura especializada.

Palabras claves: dermatitis atópica, redes neuronales profundas, procesamiento de imágenes médicas, inteligencia artificial, clasificación multiclase.

INTRODUCCIÓN

La dermatitis atópica es una enfermedad inflamatoria crónica de la piel caracterizada por resequedad, enrojecimiento, irritación y comezón persistente. Esta afección se presenta con mayor frecuencia en niños y adolescentes, aunque también puede afectar a personas adultas. Su impacto no se limita a las lesiones visibles, ya que puede alterar la

calidad de vida del paciente debido a molestias constantes, alteraciones del sueño y dificultades en actividades cotidianas.

La prevalencia de la dermatitis atópica justifica la necesidad de herramientas de detección temprana. La dermatitis atópica es la enfermedad cutánea inflamatoria más común, afecta principalmente a los niños con una prevalencia del 20% y suele iniciarse antes de los 5 años en el 90% de los casos. Se estima que esta afección impacta a entre el 15 y 20% de la población infantil, en contraste con un porcentaje mucho menor en adultos a nivel mundial (Saez-de- Ocariz y Espinosa-Padilla, 2020).

Para mitigar la subjetividad clínica, las redes neuronales convolucionales han demostrado su potencial efecto en la categorización del cáncer de piel y otras afecciones. Asimismo, esta compleja tecnología se está utilizando para identificar diferentes dermatosis, alcanzando a distinguir la mayoría de los casos que se presentan en la atención primaria de salud. En el ámbito clínico, las redes neuronales convolucionales imitan el funcionamiento de las neuronas cerebrales mediante múltiples capas interconectadas, procesando variables visuales complejas para apoyar el prediagnóstico (Guzmán-Bucio y Martínez-Morales, 2023).

El objetivo del presente trabajo es desarrollar y evaluar un sistema de clasificación de severidad de dermatitis atópica basado en redes neuronales convolucionales, como herramienta de apoyo al diagnóstico clínico en el primer nivel de atención.

MARCO TEÓRICO

La inteligencia artificial es una rama de la computación que permite desarrollar sistemas capaces de analizar información, reconocer patrones y apoyar la toma de decisiones. En el área médica, su aplicación ha crecido exponencialmente ante la necesidad de procesar grandes cantidades de datos clínicos e imágenes diagnósticas de manera rápida y objetiva.

El aprendizaje profundo, específicamente a través de las Redes Neuronales Convolucionales (CNN), ha revolucionado el procesamiento de imágenes médicas al extraer automáticamente características visuales complejas como bordes, texturas y variaciones de color. En dermatología, la literatura contemporánea demuestra que los sistemas basados en CNN pueden alcanzar niveles de precisión comparables a los de especialistas humanos. Investigaciones recientes respaldan esta eficacia; por ejemplo, Quintero (2021) validó el uso de redes convolucionales en dispositivos móviles para la identificación de enfermedades cutáneas, y Mahmoud y Salem (2024) evidenciaron su utilidad en sistemas automatizados para la detección temprana de patologías complejas como el cáncer de piel.

A pesar de estos avances tecnológicos, la implementación de modelos de aprendizaje profundo en la dermatitis atópica enfrenta un reto clínico particular: la extrema variabilidad visual de sus manifestaciones. A diferencia de afecciones con marcadores morfológicos rígidamente definidos, las lesiones atópicas son dinámicas y presentan un alto grado de solapamiento fenotípico; cambian drásticamente en función de la edad del paciente, el tono de piel, el grado de inflamación y la etapa evolutiva de la

enfermedad (Yang et al., 2020). Songjiang et al. (2021) señalan que el desarrollo de modelos predictivos basados en aprendizaje automático para esta condición requiere arquitecturas robustas capaces de generalizar sobre fenotipos superpuestos, lo cual depende críticamente de la calidad del conjunto de datos y del correcto preprocesamiento de la información.

Esta transición metodológica hacia el diagnóstico asistido por computadora plantea un profundo debate epistemológico en la práctica médica contemporánea. Autores como Martorell y Rueda Carnero (2022) exploran si la inteligencia artificial representa una amenaza o una oportunidad en la dermatología, cuestionando la validez de trasladar evaluaciones clínicas, tradicionalmente basadas en la intuición y la experiencia subjetiva humana, hacia métricas computacionales deterministas. Bajo esta visión epistemológica, el presente estudio sostiene que las redes neuronales convolucionales no tienen el propósito de sustituir el raciocinio ni la valoración médica profesional.

Por el contrario, representan una alternativa tecnológica complementaria que, al mitigar la variabilidad dependiente del observador frente al espectro continuo de la dermatitis, proporciona un marco determinista y objetivo para fundamentar decisiones clínicas más seguras y oportunas.

METODOLOGÍA

El presente estudio se fundamentó en un enfoque de investigación basado en datos secundarios, utilizando un corpus biográfico de 1,257 imágenes

dermatológicas digitales de acceso público extraídas de la plataforma Kaggle. El conjunto de datos original fue seleccionado y curado específicamente para este trabajo con el objetivo de concentrar la capacidad de generalización del modelo en la clasificación multiclase de tres estadios de severidad de la dermatitis atópica: leve ($n = 584$), moderado ($n = 293$) y grave ($n = 380$). Con el fin de optimizar la eficiencia computacional y la especificidad del clasificador en el gradiente patológico, se tomó como referencia una arquitectura inicialmente orientada a la clasificación binaria (detección de presencia o ausencia de la patología); sin embargo, esta fue modificada y adaptada para ejecutar una tarea de clasificación multiclase. De este modo, en lugar de limitar el sistema a la simple detección de si el paciente presentaba o no la afección, el enfoque se redirigió hacia la discriminación de los niveles de severidad, optimizando la capacidad del clasificador en el entorno clínico.

El repositorio consolidado fue particionado de manera aleatoria e independiente en una proporción aproximada de 80% para la fase de optimización y entrenamiento ($n = 1,005$) y un 20% retenido estrictamente como conjunto de prueba externa para la evaluación final de rendimiento ($n = 252$). La distribución cuantitativa por clases asignada al bloque de entrenamiento consistió en 467 imágenes para la clase leve, 234 para la clase moderada y 304 para la clase grave. Por su parte, el conjunto de prueba independiente conservó una distribución de 117 muestras para la clase leve, 59 muestras para la clase moderada y 76 muestras para

la clase grave, garantizando la equidad en el proceso de validación externa.

Para mitigar los efectos latentes de sobreajuste (overfitting) derivados de la dimensión acotada del dataset y mejorar la robustez del sistema ante variaciones intra-clase (tales como cambios de iluminación, ángulo u orientación), se implementó un protocolo de aumentación de datos (data augmentation) en tiempo de ejecución durante la etapa de entrenamiento. Las transformaciones estocásticas aplicadas incluyeron rotaciones aleatorias de hasta 15°, desplazamientos tanto horizontales como verticales limitados al 15%, un rango de zoom del 20% y giros horizontales (horizontal flips). Finalmente, todas las muestras del dataset fueron normalizadas y redimensionadas de forma homogénea a una resolución de 224×224 píxeles para asegurar el correcto acoplamiento con los requerimientos de la capa de entrada de la red convolucional profunda.

Arquitectura del modelo e hiperparámetros.

El sistema inteligente se fundamentó en la técnica de transferencia de aprendizaje (transfer learning) empleando la arquitectura profunda DenseNet201 preentrenada con el banco de imágenes ImageNet. Se removieron las capas completamente conectadas (top layers) originales de la red y se acopló un clasificador personalizado adaptado al problema de tres clases. La topología de este clasificador consistió jerárquicamente en: una capa de agrupación global promedio (GlobalAveragePooling2D), una capa densa intermedia de 512 neuronas con función de activación de rectificación

lineal (ReLU), una capa de regularización Dropout con una tasa del 0.5 para prevenir la coadaptación de características, y finalmente, una capa de salida densa con 3 neuronas y activación Softmax para la distribución de probabilidad multiclase.

El entrenamiento se gestionó mediante el optimizador Adam con una tasa de aprendizaje (learning rate) inicial estática de 0.0001 y la función de pérdida de entropía cruzada categórica (categorical crossentropy). El proceso se configuró para un máximo de 100 épocas con un tamaño de lote (batch size) de 40 imágenes. Asimismo, se incorporaron dos funciones de retrollamada (callbacks) dinámicas para optimizar la convergencia: ReduceLROnPlateau, programado para reducir el learning rate por un factor de 0.2 ante una meseta en la pérdida de validación con una paciencia de 5 épocas (límite mínimo de 1×10^{-6}), y EarlyStopping, parametrizado con una paciencia de 10 épocas para interrumpir prematuramente el entrenamiento si cesaban las mejoras en el conjunto de validación.

Como se observa en las curvas de rendimiento empíricas, este mecanismo interrumpió el proceso de optimización en la época 20 al estabilizarse la pérdida de validación. El desempeño global y por categoría del modelo final fue evaluado formalmente mediante las métricas de exactitud (accuracy), matriz de confusión, sensibilidad (recall), especificidad y el área bajo la curva de característica operativa del receptor macro (AUC-ROC macro).

RESULTADOS DEL ESTUDIO

El proceso de optimización y el análisis del rendimiento del sistema inteligente NEURODERM se ejecutaron de manera estructurada a través de las curvas de aprendizaje y la evaluación de las matrices de confusión frente al conjunto de prueba externa.

El comportamiento iterativo del entrenamiento y la validación del modelo basado en la arquitectura DenseNet201 se ilustró a través del comportamiento de la pérdida y la exactitud a lo largo de las épocas (Figura 1). La función de pérdida (Cross Entropy Loss) exhibió un descenso continuo desde la época inicial, donde el conjunto de entrenamiento (train) partió de un valor aproximado de 1.07 y el de validación (val) de 0.77. Ambos trazados convergieron de manera decreciente y asintótica hacia la época 20, estabilizando la pérdida de validación en un rango cercano a 0.58.

De manera homóloga, la exactitud de la clasificación (Classification Accuracy) reflejó una progresión ascendente estable. El conjunto de validación mostró una respuesta favorable desde las primeras interacciones, estabilizándose a partir de la época 10 y alcanzando su meseta de rendimiento final entre el 74% y el 75% en las épocas de cierre. La ausencia de divergencias o incrementos tardíos en la curva de pérdida de validación confirmó la efectividad de las funciones de retrollamada (callbacks) implementadas para detener el proceso de optimización de forma oportuna, eludiendo comportamientos asociados al sobreajuste (overfitting).

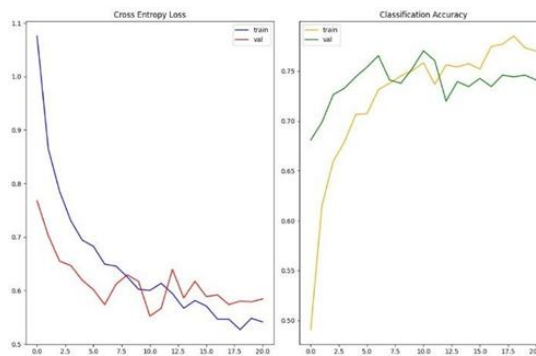


Figura 1. Curvas de pérdida categórica (Cross Entropy Loss) y exactitud (Classification Accuracy) durante las 20 épocas de entrenamiento y validación del modelo DenseNet201.

Evaluación Global y Categórica del Rendimiento Una vez concluida la etapa de optimización, el clasificador final se evaluó ante el conjunto de prueba independiente ($n = 252$). A nivel global, el sistema inteligente alcanzó una exactitud (accuracy) del 77.04% y un AUC-ROC macro del 91.34%, demostrando una sólida capacidad general de discriminación estadística en una tarea de tres clases.

Para analizar la distribución de los aciertos y las desviaciones del modelo en cada estadio patológico, se examinaron los valores cuantitativos absolutos registrados en la matriz de confusión (Figura 2). De las 252 muestras totales evaluadas, el clasificador identificó correctamente a 234 imágenes pertenecientes a la clase leve, 82 a la clase moderada y 157 a la clase grave, consolidando un total de 473 predicciones verdaderas positivas.

Evaluación Global y Categórica del Rendimiento

Una vez concluida la etapa de optimización, el clasificador final se evaluó ante el conjunto de prueba independiente ($n = 279$). A nivel global, el sistema inteligente alcanzó una exactitud

(accuracy) del 84.95% y un AUC-ROC macro del 94.22%, demostrando una sólida capacidad general de discriminación estadística en una tarea de tres clases.

Para analizar la distribución de los aciertos y las desviaciones del modelo en cada estadio patológico, se examinaron los valores cuantitativos absolutos registrados en la matriz de confusión (Figura 2). De las 279 muestras totales evaluadas, el clasificador identificó correctamente a 119 imágenes pertenecientes a la clase leve, 31 a la clase moderada y 91 a la clase grave, consolidando un total de 241 predicciones verdaderas positivas.

A partir de las frecuencias absolutas, se calcularon y analizaron las métricas normalizadas de sensibilidad (recall) y especificidad por cada categoría clínica (Tabla 1).

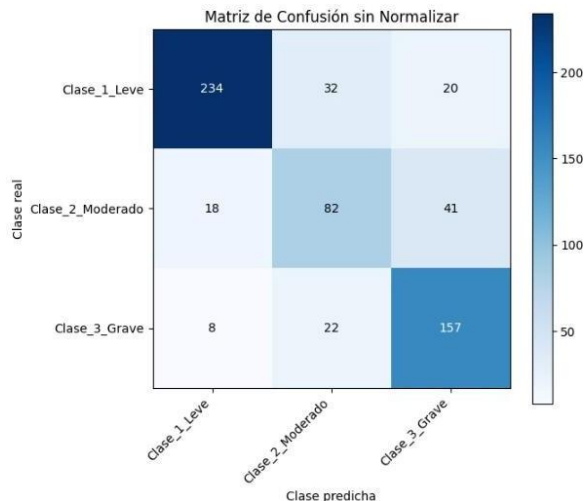


Figura 2. Matriz de confusión en valores absolutos para el conjunto de prueba externa de 252 imágenes, evaluando las clases Leve, Moderada y Grave.

Clase de Severidad	Sensibilidad (Recall)	Especificidad	Prevalencia en Prueba
Clase 1: Leve	81.82%	92.07%	46.58%
Clase 2: Moderado	58.16%	88.58%	22.96%
Clase 3: Grave	83.96%	85.71%	30.46%

El desglose de rendimiento evidenció que los valores más altos de sensibilidad se concentraron en los extremos del gradiente clínico. La clase grave obtuvo el índice de reconocimiento correcto más elevado con un 83.96% de sensibilidad y una especificidad del 85.71%. Por su parte, la clase leve registró una sensibilidad del 81.82%, acompañada de la especificidad más alta obtenida en el estudio, situándose en un 92.07%.

En lo que respecta a la categoría intermedia (moderada), el sistema obtuvo una sensibilidad del 58.16% y una especificidad del 88.58%. La distribución porcentual de los errores de clasificación para esta etapa de transición se detalló en la matriz de confusión normalizada (Figura 3). Las desviaciones de la clase moderada se desplazaron predominantemente hacia las categorías contiguas: un 12.77% de las muestras reales moderadas fueron clasificadas erróneamente como leves, mientras que un 29.08% se asignaron de forma incorrecta a la clase grave. En contraste, las transiciones erróneas directas entre los extremos clínicos fueron mínimas, registrando únicamente un 6.99% de imágenes leves predichas como graves y un 4.28% de imágenes graves etiquetadas como leves.

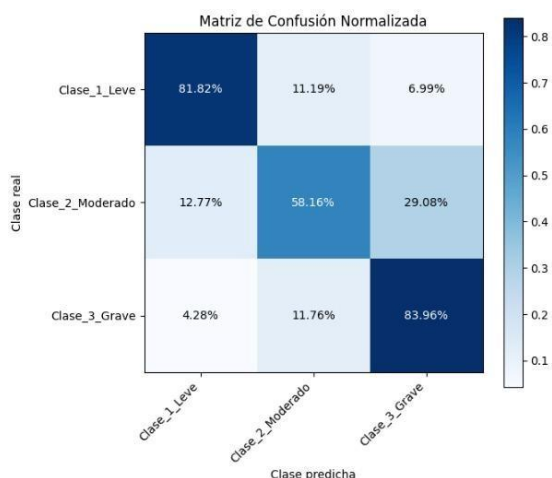


Figura 3. Matriz de confusión normalizada reflejando la sensibilidad (recall) alcanzada de forma independiente para cada nivel de severidad en el conjunto de prueba.

DISCUSIÓN (O ANÁLISIS DE RESULTADOS)

Los hallazgos del estudio respaldan la hipótesis de que un sistema basado en aprendizaje profundo puede actuar como una herramienta efectiva para mitigar la subjetividad en la evaluación visual de la dermatitis atópica. El rendimiento obtenido en la identificación de la severidad leve (81.82% de sensibilidad) responde al objetivo principal de la investigación: proveer un soporte diagnóstico temprano para el primer nivel de atención médica, facilitando la intervención oportuna en contextos con recursos sanitarios limitados.

La diferencia de rendimiento observada en la categoría moderada (58.16% de acierto) constituye un fenómeno técnico y clínico congruente con los desafíos reportados en el análisis de imágenes dermatológicas. Autores como Guzmán-Bucio y Martínez-Morales (2023) y Quintero (2021) señalan que la alta variabilidad visual de las manifestaciones cutáneas dificulta la delimitación estricta entre clases.

Clínicamente, la dermatitis atópica moderada actúa como una fase de transición visual continua donde la intensidad del eritema, la descamación y la irritación pueden solaparse con las etapas contiguas. Esto explica el comportamiento del modelo en el conjunto de prueba, donde de las 59 imágenes reales de la categoría moderada, la red convolucional interpretó erróneamente 8 muestras como leves y 17 como graves, reflejando de forma fidedigna la naturaleza difusa de las fronteras fenotípicas de la enfermedad.

A pesar de la solapación visual en la etapa intermedia, la estabilidad observada en las curvas de pérdida y exactitud demuestra que la arquitectura implementada extrajo de manera óptima los patrones visuales clave sin caer en el sobreajuste. Estos resultados representan un punto de partida prometedor, si bien se requiere ampliar el conjunto de datos y refinar el manejo de la clase moderada antes de considerar su implementación clínica.

CONCLUSIÓN

El desarrollo del sistema NEURODERM, mediante transferencia de aprendizaje con la arquitectura DenseNet201, representa una aproximación tecnológica viable para apoyar el diagnóstico asistido de la dermatitis atópica. El sistema demostró una capacidad sólida para discriminar los estadios extremos de la patología, alcanzando sensibilidades del 81.82% y 83.96% para las clases leve y grave, respectivamente. Esto evidencia su potencial como herramienta auxiliar en el primer nivel de atención médica, donde puede ofrecer un criterio de respaldo automatizado frente a la variabilidad

clínica del padecimiento. Sin embargo, el balance final de la investigación indica que el modelo requiere mayor madurez antes de mitigar de forma integral la subjetividad clínica, dado su rendimiento global del 77.04% de exactitud. La principal limitación de este estudio radica en la dificultad del sistema para clasificar la categoría moderada, la cual obtuvo un 58.16% de sensibilidad. Este déficit refleja el profundo reto técnico de modelar computacionalmente las fronteras fenotípicas difusas y la transición visual continua propias de esta enfermedad. Adicionalmente, el entrenamiento del modelo dependió de un corpus acotado de 1,257 imágenes secundarias de acceso público, lo que restringe su capacidad de generalización frente a la amplia diversidad demográfica y clínica de los pacientes en un entorno real.

Como líneas de trabajo futuro, resulta imperativo ampliar y balancear el conjunto de datos integrando muestras clínicas primarias que representen de manera más fiel la diversidad de los tonos de piel y las condiciones de iluminación del entorno médico local. Asimismo, las futuras investigaciones deberán explorar enfoques metodológicos complementarios, como la integración de metadatos clínicos del paciente o el uso de métodos de ensamble, para robustecer la extracción de características en las fases de transición de la dermatitis. Estas iteraciones desde una prueba de concepto hacia una herramienta clínicamente implementable que logre democratizar el acceso a prediagnósticos confiables en contextos con infraestructura limitada.

REFERENCIAS

- Guzmán-Bucio, S., & Martínez-Morales, V. (2023). Inteligencia artificial en dermatología. *Dermatología Revista Mexicana*, 67(6), 905 – 910.
- Mahmoud, N. M., & Salem, A. M. (2024). Early automated detection system for skin cancer diagnosis using artificial intelligent techniques. *Scientific Reports*, 14(1).
- Martorell, A., & Rueda Carnero, J. M. (2022). Inteligencia artificial en dermatología: ¿Amenaza u oportunidad? *Actas Dermo-Sifiliográficas*, 113(1), 30–46.
- Quintero, G. J. (2021). Uso de redes neuronales convolucionales en teléfonos inteligentes para la identificación de enfermedades cutáneas. *IDEULA*, 81–98.
- Saez-de-Ocariz, M. y Espinosa-Padilla, S. E. (2020). Atención y tratamiento integral de la dermatitis atópica. *Acta Pediátrica de México*, 41, 286-292.
- Songjiang, W., et al. (2021). Machine learning-based prediction models for atopic dermatitis. *Computational and Structural Biotechnology Journal*, 19, 3737–3745.
- Yang, Y. B., et al. (2020). Common atopic dermatitis rating scales: A practical approach and brief review. *Journal of Cutaneous Medicine and Surgery*, 24(4), 399–400.