PRISMA ODS
REVISTA MULTIDISCIPLINARIA
SOBRE DESARROLLO SOSTENIBLE
ISSN: 3072-8452

INTELIGENCIA ARTIFICIAL CONTRA INTELIGENCIA ARTIFICIAL EN LA DETECCIÓN DE DEEPFAKES MEDIANTE UNA ARQUITECTURA HÍBRIDA DE REDES NEURONALES PROFUNDAS

ARTIFICIAL INTELLIGENCE VERSUS ARTIFICIAL INTELLIGENCE IN DEEPFAKE DETECTION USING A HYBRID DEEP NEURAL NETWORK ARCHITECTURE

AUTORES

DANNY FABIAN CANDO GORDON
UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI
ECUADOR

GUILLERMO FERNANDO ARGOTI SANTACRUZ
UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI
ECUADOR

MARCO ANTONIO YANDÚN VELASTEGUÍ
UNIVERSIDAD POLITÉCNICA ESTATAL DEL CARCHI
ECUADOR

doi 10.65011

Volumen 5 - Número 6
2026

Redilat
Red de Investigadores Latinoamericanos
PRISMA ODS
REVISTA MULTIDISCIPLINARIA SOBRE DESARROLLO SOSTENIBLE
LATINDES
INSTITUTO LATINOAMERICANO EN DESARROLLO SOSTENIBLE

Inteligencia Artificial contra Inteligencia Artificial en la Detección de Deepfakes mediante una Arquitectura Híbrida de Redes Neuronales Profundas

Artificial Intelligence versus Artificial Intelligence in Deepfake Detection Using a Hybrid Deep Neural Network Architecture

Danny Fabian Cando Gordon
danny.cando@upec.edu.ec
https://orcid.org/0009-0008-4559-5657
Universidad Politécnica Estatal del Carchi
Tulcán – Ecuador

Guillermo Fernando Argoti Santacruz
guillermo.argoti@upec.edu.ec
https://orcid.org/0009-0007-4735-9589
Universidad Politécnica Estatal del Carchi
Tulcán – Ecuador

Marco Antonio Yandún Velasteguí
marco.yandun@upec.edu.ec
https://orcid.org/0000-0001-5627-9838
Universidad Politécnica Estatal del Carchi
Tulcán – Ecuador

Artículo recibido: 23/10/2026
Aceptado para publicación: 23/11/2026
Conflictos de Intereses: Ninguno que declarar.

RESUMEN

Un detector de deepfakes entrenado con las imágenes de un generador suele reconocerlas casi sin errores. El problema aparece cuando surge un modelo nuevo, cuando la imagen se comprime a la circular en redes o cuando alguien la altera para engañar al sistema; en esas condiciones, los detectores que dependen de una sola señal fallan mucho más de lo que sugieren sus cifras de laboratorio. Frente a esta limitación planteamos ESPEJO, una arquitectura de software en la que la inteligencia artificial revisa lo que ella misma es capaz de producir. El trabajo se apoya en la investigación basada en diseño y parte de la revisión crítica de 30 estudios revisados por pares, publicados entre 2022 y 2024, que elegimos porque tratan de forma directa la generalización y la robustez de los detectores. De esa revisión salieron nueve requisitos, y con ellos definimos siete módulos. Cada imagen pasa por tres análisis que no dependen entre sí. El primero usa lo que el modelo CLIP ya conoce de las fotografías reales; el segundo busca los patrones repetitivos que suele dejar el sobremuestreo; el tercero observa qué tan difícil le resulta al autocodificador de un modelo de difusión reconstruir la imagen. A esto se suma un verificador de marcas de agua, y un metaclasificador calibrado reúne toda la evidencia. Cuando la información no alcanza, el sistema puede dejar el caso como indeterminado en lugar de forzar una respuesta. Nuestra conclusión es que un solo detector rara vez logra generalizar bien; eso se consigue al combinar enfoques diferentes y al mostrar con claridad cuándo el sistema no está seguro. El artículo incluye también un protocolo de siete experimentos, pensado para que otros investigadores puedan replicarlo y poner a prueba la arquitectura.

Palabras clave: deepfakes, redes neuronales profundas, modelos de difusión, visión por computadora, análisis forense digital

ABSTRACT

A deepfake detector trained on images from one generator usually recognizes them almost flawlessly. Trouble begins when a new model appears, when the image is compressed as it circulates online, or when someone alters it to fool the system; under those conditions, detectors that rely on a single signal fail far more often than their lab figures suggest. To address this weakness we propose ESPEJO, a software architecture that sets artificial intelligence to watch over itself. We followed design science research and started from a critical review of 30 peer-reviewed studies published between 2022 and 2024, chosen for their contribution to the generalization and robustness of detection. That analysis yielded nine requirements that shaped seven modules. Each image is examined along three independent paths: one draws on what the CLIP model knows about real photographs, another tracks the regularities left by upsampling, and the third measures how hard it is for the autoencoder of a diffusion model to rebuild the image. A watermark verifier completes the evidence, and a calibrated meta-classifier integrates it, with the option of labeling a case as undetermined when the data fall short. We conclude that generalization is unlikely to come from an isolated detector: it has to be built by combining different principles and making uncertainty visible. The article includes a reproducible seven-experiment protocol to test the proposal.

Keywords: deepfakes, deep neural networks, diffusion models, computer vision, digital forensics

INTRODUCCIÓN

Hace no tanto, una fotografía bastaba como prueba. Se podía retocar o sacar de contexto, pero inventar de cero una escena creíble costaba tiempo y oficio. Eso cambió en pocos años. Nightingale y Farid (2022) pidieron a cientos de personas que separaran rostros reales de rostros creados por redes adversarias y comprobaron que no lo lograban; peor aún, los rostros falsos les inspiraban, en promedio, 7,7 % más confianza. Los estudios posteriores, ya con imágenes de modelos de difusión y participantes de varios países, llegaron a la misma conclusión (Frank et al., 2024; Lu et al., 2023). Buena parte del cambio se explica por los modelos de difusión latente, que trasladaron la generación a un espacio comprimido y la hicieron posible en un computador corriente (Rombach et al., 2022).

El asunto rebasa lo técnico. Cuando cualquier imagen puede ser fabricada, se resiente el derecho de las personas a recibir información veraz y se debilita la confianza en la prensa, en los tribunales y en las instituciones que dependen de la evidencia visual. Son preocupaciones que recoge el Objetivo de Desarrollo Sostenible 16, en particular su meta 16.10 sobre el acceso público a la información. Contar con herramientas abiertas y verificables para rastrear de dónde proviene una imagen también tiene que ver con el ODS 9, que promueve la innovación y el fortalecimiento de capacidades tecnológicas propias en regiones como América Latina.

Lo paradójico es que la mejor defensa viene de la misma tecnología. Hablamos de redes neuronales profundas entrenadas para reconocer lo que producen otras redes. Las revisiones más recientes señalan que hoy este es el enfoque dominante, y también coinciden en su punto débil, que es la generalización (Heidari et al., 2024; Passos et al., 2024; Rana et al., 2022). Un detector diseñado para redes adversarias rinde bastante peor cuando se enfrenta a imágenes de difusión (Corvi et al., 2023; Ricker, Damm, et al., 2024). Ojha et al. (2023) explicaron por qué ocurre: el clasificador se queda con las huellas del generador que usó para entrenarse y, cuando algo no encaja con lo aprendido, lo toma como real. Con generadores nuevos apareciendo cada poco mes, ese punto ciego deja de ser una excepción y se vuelve lo normal (Epstein et al., 2023).

Los primeros detectores optaron por el camino más directo, que fue entrenar una red con ejemplos etiquetados de imágenes reales y falsas. En el caso de los rostros aparecieron propuestas bastante creativas. Shiohara y Yamasaki (2022) combinaron cada foto con una copia ligeramente modificada de sí misma para que la red aprendiera a detectar costuras de

tipo general. Haliassos et al. (2022) analizaron si el movimiento de los labios coincidía con la voz en videos reales, y Coccomini et al. (2022) integraron una EfficientNet con un Vision Transformer. Después, Tan et al. (2023) entrenaron su modelo a partir de los gradientes de una red preentrenada, y Yan, Zhang, Fan, et al. (2023) buscaron conservar únicamente los rasgos comunes a todas las falsificaciones. El trabajo de Bird y Lotfi (2024) dejó un hallazgo llamativo. Cuando examinaron en qué partes de la imagen se fijaba su red para decidir, encontraron que miraba sobre todo los defectos del fondo y casi nunca el objeto principal.

El cambio importante vino con los modelos fundacionales. Ojha et al. (2023) ya no entrenaron un detector, sino que clasificaron las imágenes dentro del espacio de rasgos de CLIP. Con un método así de simple superaron al estado del arte por 15,07 puntos de mAP ante generadores que el sistema no había visto. Cozzolino, Poggi, Corvi, et al. (2024) comprobaron que unas pocas imágenes de un solo generador alcanzan para identificar las de herramientas comerciales, y Liu et al. (2024) adaptaron esas representaciones al análisis forense. Hay otra línea de trabajo, quizá la más curiosa, que usa al propio generador como herramienta de prueba. Wang et al. (2023) observaron que un modelo de difusión reconstruye con facilidad lo que él mismo pudo haber generado, mientras que una foto real le resulta más difícil. Ricker, Lukovnikov, et al. (2024) aplicaron esta idea al autocodificador latente y eliminaron la necesidad de entrenamiento. Chen et al. (2024) usaron la reconstrucción para crear ejemplos difíciles, y Cozzolino, Poggi, Nießner, et al. (2024) propusieron una detección de disparo cero. De forma paralela, Tan et al. (2024) identificaron en el sobremuestreo una huella estructural que se repetía en 28 generadores diferentes.

Hay también una alternativa preventiva. Consiste en marcar las imágenes desde el momento en que se generan (Fernandez et al., 2023; Wen et al., 2023) o en reconocer qué modelo las creó (Sha et al., 2023). Saberi et al. (2024) mostraron que tanto las marcas como los clasificadores tienen límites teóricos que ningún ajuste logra eliminar por completo. Los bancos de prueba como GenImage (Zhu et al., 2023), DeepfakeBench (Yan, Zhang, Yuan, et al., 2023) y el protocolo en línea de Epstein et al. (2023) llegaron, cada uno por su lado, al mismo resultado: el rendimiento baja apenas se cambia de generador.

Lo que encontramos en la literatura son soluciones parciales muy útiles, aunque casi siempre trabajan por separado. Cada método se basa en un único principio y carga con sus propias debilidades. Nuestra propuesta parte de una idea sencilla. Si esos principios fallan por motivos diferentes, combinarlos de forma ordenada debería hacer que el sistema sea más

difícil de engañar. Es posible que un generador nuevo logre evadir una de las señales, pero es poco probable que evada al mismo tiempo las representaciones semánticas, las huellas del sobremuestreo y la firma de su propio autocodificador.

Esto nos lleva a la pregunta que orienta el estudio: ¿qué requisitos debe reunir y cómo debe estar organizada una arquitectura de detección para seguir siendo eficaz ante generadores que no conoció durante el entrenamiento? Como respuesta provisional proponemos tres hipótesis que la arquitectura permite poner a prueba. **H1:** con imágenes de generadores que no se incluyeron en el entrenamiento, la fusión de varias ramas obtiene mejores resultados que la mejor rama por sí sola. **H2:** cuando las imágenes se comprimen o cambian de tamaño, la fusión pierde menos rendimiento que cada rama por separado. **H3:** una zona de abstención disminuye los errores en las decisiones que el sistema sí toma, a cambio de dejar sin responder un porcentaje controlado de casos.

El artículo hace tres aportes. El primero es sintetizar en nueve requisitos lo que muestran 30 estudios publicados entre 2022 y 2024. El segundo es presentar ESPEJO, una arquitectura modular que integra varias ramas de detección, una fusión calibrada que puede abstenerse y un módulo que explica cada decisión. El tercero es proponer un protocolo de evaluación reproducible para comprobar si la arquitectura cumple lo que plantea. El nombre refleja esa intención, que es usar a la inteligencia artificial como un espejo de sí misma.

METODOLOGÍA

Diseño de la investigación

Optamos por la investigación basada en diseño, un enfoque frecuente en la ingeniería de software. Aquí el producto científico es un artefacto, en este caso una arquitectura, que tiene que resolver un problema real y traer consigo una manera de evaluarlo. El estudio es cualitativo y de alcance propositivo; en esta etapa no es experimental, porque diseñamos la arquitectura y dejamos especificado el protocolo con el que podrá ponerse a prueba. El trabajo avanzó en cuatro fases. Primero delimitamos el problema a partir de la evidencia sobre la pérdida de generalización de los detectores; luego definimos los requisitos mediante la revisión crítica de la literatura; después tradujimos cada requisito en un componente y, por último, especificamos cómo evaluar el conjunto.

Unidad de análisis

El estudio no involucró participantes humanos. Las unidades de análisis fueron 30 estudios revisados por pares: 6 publicados en 2022, 12 en 2023 y 12 en 2024. Proceden de las conferencias de mayor impacto en visión por computadora y aprendizaje automático (CVPR, ICCV, ECCV, NeurIPS, ICML e ICLR), de conferencias de seguridad y procesamiento de señales (IEEE S&P, ACM CCS, ICASSP y VISAPP) y de revistas indexadas en Scopus y Web of Science. Abarcan métodos de detección, bancos de evaluación, estudios sobre percepción humana, trabajos sobre marcas de agua y revisiones sistemáticas.

Procedimiento de selección

La selección fue intencional y guiada por criterios. Consultamos IEEE Xplore, CVF Open Access, OpenReview, PMLR y Google Scholar con la cadena ("deepfake detection" OR "AI-generated image detection" OR "synthetic image detection") AND ("diffusion model" OR "GAN") AND ("generalization" OR "robustness"), y completamos la búsqueda revisando las referencias de los trabajos más citados. La Tabla 1 resume los criterios aplicados.

Tabla 1. Criterios de inclusión y exclusión de los estudios

CRITERIO	*INCLUSIÓN*	*EXCLUSIÓN*
Periodo	Publicaciones de 2022 a 2024	Publicaciones anteriores a 2022
Tipo de publicación	Revisión por pares en conferencia o revista	Preprints, tesis y blogs
Tema	Detección de imágenes generadas o manipuladas por IA; percepción humana; límites de la detección	Trabajos centrados solo en audio o texto
Evidencia	Evaluación experimental reproducible o revisión sistemática	Textos de opinión sin datos

Fuente: Elaboración propia.

Técnicas de recolección y análisis de datos

Para cada estudio registramos en una matriz de extracción el medio de publicación, la arquitectura empleada, los generadores evaluados, el protocolo de generalización, el principio de detección, su punto ciego y los resultados tal como los reportaron sus autores. No recalculamos métricas, de modo que las cifras citadas deben leerse dentro del protocolo de cada estudio. Los trabajos se agruparon de forma inductiva según su principio de detección, y las conclusiones de cada grupo se contrastaron con los bancos de evaluación independientes

del corpus. Para que cada decisión de diseño pueda rastrearse, todo requisito quedó vinculado al menos a un estudio que lo fundamenta (Tabla 3).

RESULTADOS Y DISCUSIÓN

Los resultados se presentan en cuatro partes: los puntos ciegos que deja la literatura, los requisitos que se desprenden de ellos, la arquitectura que los atiende y el protocolo con el que puede ponerse a prueba. Al final discutimos su alcance.

Puntos ciegos de la literatura y su lugar en la arquitectura

El análisis del corpus mostró que cada línea de detección resuelve un problema y deja otro abierto. Las representaciones de CLIP generalizan bien, pero dependen de un modelo ajeno; las huellas del sobremuestreo se degradan con la compresión; la reconstrucción funciona mientras los generadores sigan usando difusión latente, y las marcas de agua solo sirven si el desarrollador decidió incluirlas. La Tabla 2 recoge estas líneas y el componente de ESPEJO que incorpora cada una.

Tabla 2. Líneas de detección, punto ciego principal y componente de ESPEJO que las incorpora

LÍNEA	*ESTUDIOS BASE*	*PUNTO CIEGO PRINCIPAL*	*COMPONENTE*
Modelos fundacionales	Ojha et al. (2023); Cozzolino, Poggi, Corvi, et al. (2024); Liu et al. (2024)	Dependencia de un modelo externo	Rama semántica (B1)
Artefactos estructurales	Tan et al. (2024); Tan et al. (2023)	Sensibles al posprocesamiento	Rama estructural (B2)
Reconstrucción	Ricker, Lukovnikov, et al. (2024); Wang et al. (2023); Chen et al. (2024)	Ligada a la difusión latente	Rama de reconstrucción (B3)
Marcas de agua y atribución	Fernandez et al. (2023); Wen et al. (2023); Sha et al. (2023)	Requiere cooperación del desarrollador	Módulo de procedencia (M4)
Clasificadores explicables	Bird y Lotfi (2024); Yan, Zhang, Fan, et al. (2023)	Memorizan huellas del generador	Módulo de explicabilidad (M6)
Bancos y límites teóricos	Zhu et al. (2023); Epstein et al. (2023); Saberi et al. (2024)	Protocolos heterogéneos	Protocolo de evaluación

Fuente: Elaboración propia a partir de los estudios revisados.

Requisitos de diseño

De esos puntos ciegos salieron cinco requisitos funcionales y cuatro no funcionales (Tabla 3). Los funcionales describen lo que el sistema debe hacer: reconocer imágenes de generadores que nunca vio, combinar señales independientes, verificar marcas de agua, abstenerse cuando la evidencia no alcance y explicar cada decisión. Los no funcionales fijan cómo debe hacerlo: con robustez frente a la compresión, con probabilidades calibradas, de forma modular y reproducible.

Tabla 3. Requisitos de la arquitectura y su fundamento en la literatura

CÓDIGO	***REQUISITO***	***TIPO***	***FUNDAMENTO***
RF1	Detectar imágenes de generadores no vistos en el entrenamiento	Funcional	Ojha et al. (2023); Epstein et al. (2023)
RF2	Combinar principios de detección independientes	Funcional	Tan et al. (2024); Ricker, Lukovnikov, et al. (2024)
RF3	Verificar marcas de agua y metadatos cuando existan	Funcional	Fernandez et al. (2023); Wen et al. (2023)
RF4	Abstenerse cuando la evidencia sea insuficiente	Funcional	Saberi et al. (2024)
RF5	Explicar la decisión con evidencia visual	Funcional	Bird y Lotfi (2024)
RNF1	Robustez ante compresión y redimensionamiento	No funcional	Corvi et al. (2023); Zhu et al. (2023)
RNF2	Probabilidades calibradas, no solo etiquetas	No funcional	Saberi et al. (2024)
RNF3	Modularidad para incorporar nuevas ramas	No funcional	Epstein et al. (2023)
RNF4	Reproducibilidad del entrenamiento y la evaluación	No funcional	Yan, Zhang, Yuan, et al. (2023)

Fuente: Elaboración propia. RF = requisito funcional; RNF = requisito no funcional.

Arquitectura ESPEJO

ESPEJO consta de siete módulos repartidos en cuatro capas: ingesta, detección en paralelo, fusión y servicio (Figura 1). Procesar en paralelo fue una decisión deliberada que responde al requisito RF2. Cada rama aplica un principio distinto y no depende de lo que digan las demás, así que si una falla no arrastra al resto.

Figura 1. Arquitectura modular de ESPEJO



Fuente: Elaboración propia. M5, el módulo resaltado, es el que toma la decisión final. Las ramas B1, B2 y B3 trabajan como detectores pasivos, y M4 cumple la función de verificador preventivo.

M0. Ingesta y preprocesamiento. Recibe la imagen y conserva una copia sin tocar para M4. Con ella prepara recortes a distintas escalas, sin cambiar el tamaño del original, porque la interpolación elimina justo las huellas de alta frecuencia de las que depende B2 (Corvi et al., 2023; Ricker, Damm, et al., 2024).

B1. Rama semántica. Extrae el vector de rasgos de un codificador CLIP congelado y lo clasifica mediante una sonda lineal entrenada con pocas imágenes, siguiendo la propuesta de Ojha et al. (2023) y Cozzolino, Poggi, Corvi, et al. (2024). Decidimos no ajustar el codificador para conservar toda la amplitud de lo que ya aprendió.

B2. Rama estructural. Obtiene las relaciones entre píxeles vecinos, que son las que revelan el sobremuestreo, y las procesa con una red residual ligera, como hacen Tan et al. (2024). Esta rama se fija sobre todo en los detalles locales y es también la que menos recursos de cómputo necesita.

B3. Rama de reconstrucción. Hace pasar la imagen por el autocodificador de un modelo de difusión latente y calcula la distancia perceptual entre la imagen original y la reconstruida, de forma similar a AEROBLADE (Ricker, Lukovnikov, et al., 2024). En esta rama se concreta

la idea que da nombre al artículo: el mismo generador termina siendo la herramienta que lo descubre. Pueden usarse varios autocodificadores en paralelo, uno para cada familia de modelos.

M4. Procedencia. Revisa si la imagen tiene marcas de agua de los esquemas conocidos (Fernandez et al., 2023; Wen et al., 2023) y lee sus metadatos de origen. El resultado es uno de tres: marca presente, marca ausente o no verificable. Hallar una marca válida es una prueba sólida, pero no hallarla no significa que la imagen sea auténtica (Saberi et al., 2024).

M5. Fusión calibrada. Toma las tres puntuaciones junto con la señal de procedencia y las une mediante una regresión logística, entrenada con un conjunto de validación diferente al que usaron las ramas. Las probabilidades se calibran con escalado de temperatura (RNF2). Si el resultado queda dentro de una franja de incertidumbre, el sistema prefiere abstenerse antes que dar un veredicto forzado (RF4).

M6. Explicabilidad y servicio. Genera un mapa de activación para B2 y el mapa de error de reconstrucción de B3, este último útil para ubicar zonas editadas (Ricker, Lukovnikov, et al., 2024). Los resultados se entregan mediante una API REST.

La Figura 2 resume el procedimiento central de inferencia.

Figura 2. Algoritmo de inferencia de ESPEJO

Algoritmo 1. Inferencia en ESPEJO
Entrada: imagen x; umbrales τ_bajo, τ_alto; temperatura T
Salida: etiqueta ∈ {REAL, SINTÉTICA, INDETERMINADA}, probabilidad p, evidencia E

```
x₀, R ← Preprocesar(x)              # original y recortes multiescala
w ← VerificarProcedencia(x₀)        # {presente, ausente, no_verificable}
si w = presente entonces
    devolver SINTÉTICA, 1.0, {marca de agua}
s₁ ← SondaLineal(CLIP(R))           # rama semántica B1
s₂ ← ResNetLigera(NPR(R))           # rama estructural B2
s₃ ← máx_k LPIPS(x₀, AE_k(x₀))      # rama de reconstrucción B3
z ← MetaLogística(s₁, s₂, s₃, w)
p ← σ(z / T)                        # calibración por temperatura
si τ_bajo < p < τ_alto entonces
    etiqueta ← INDETERMINADA
sino etiqueta ← SINTÉTICA si p ≥ τ_alto; REAL en otro caso
E ← {GradCAM(B2), MapaError(B3), (s₁, s₂, s₃)}
devolver etiqueta, p, E
```

Fuente: Elaboración propia.

En el paso 7 se elige el valor más informativo de entre varios autocodificadores, con el fin de cubrir distintas familias de difusión. Los umbrales τ_bajo y τ_alto se fijan en validación de acuerdo con el costo que tenga un error en cada contexto, ya que un falso positivo no pesa lo mismo en una sala de redacción que en un peritaje judicial. La Tabla 4 recoge las tecnologías previstas para cada componente.

Tabla 4. Stack tecnológico previsto para ESPEJO

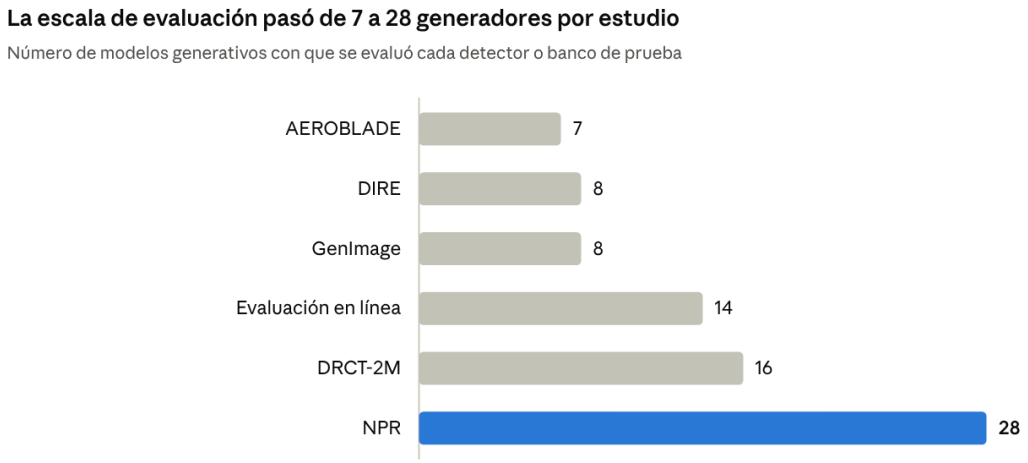
CAPA	*COMPONENTE*	*TECNOLOGÍA*	*JUSTIFICACIÓN*
Ingesta	M0	Python, Pillow, OpenCV	Lectura sin pérdida y recortes multiescala
Detección	B1	PyTorch, OpenCLIP (ViT-L/14)	Codificador preentrenado de uso abierto
Detección	B2	PyTorch, ResNet ligera	Bajo costo e inferencia rápida
Detección	B3	Diffusers (autocodificadores de difusión latente), LPIPS	Reconstrucción sin entrenamiento adicional
Procedencia	M4	Detectores de marca de agua de referencia; lector de metadatos	Verificación preventiva
Fusión	M5	scikit-learn (regresión logística, calibración)	Modelo simple, interpretable y auditable
Servicio	M6	FastAPI, Grad-CAM, Docker	Despliegue reproducible como microservicio

Fuente: Elaboración propia. Las versiones concretas se fijan en un archivo de dependencias para garantizar la reproducibilidad (RNF4).

Protocolo de evaluación

Una arquitectura solo aporta conocimiento si alguien puede intentar refutarla. Por eso detallamos un protocolo que cualquier grupo podría ejecutar con recursos públicos y que pone a ESPEJO justo en las situaciones donde los detectores suelen tropezar. Como referencia de lo que hoy se exige, la escala de evaluación en la literatura pasó de 7 a 28 generadores por estudio (Figura 3).

Figura 3. Número de generadores con que se evaluó cada detector o banco de prueba



Fuente: Elaboración propia con datos reportados por Ricker, Lukovnikov, et al. (2024), Wang et al. (2023), Zhu et al. (2023), Epstein et al. (2023), Chen et al. (2024) y Tan et al. (2024).

La Tabla 5 describe los siete experimentos previstos y su relación con las hipótesis y los requisitos.

Tabla 5. Diseño experimental para validar la arquitectura

EXPERIMENTO	*PROPÓSITO*	*DATOS*	*MÉTRICAS*
E1. Exclusión de generadores	Contrastar H1 (RF1)	GenImage (Zhu et al., 2023); DRCT-2M (Chen et al., 2024)	AUC; precisión media (AP)
E2. Evaluación en línea	Simular la llegada de nuevos generadores (RNF3)	Orden de publicación, según Epstein et al. (2023)	AUC en cada paso
E3. Robustez	Contrastar H2 (RNF1)	Compresión JPEG, redimensionamiento y desenfoque	Variación de AUC respecto a imágenes limpias
E4. Ablación	Medir el aporte de cada rama (RF2)	Mismo conjunto que E1, retirando una rama cada vez	Diferencia de AUC
E5. Calibración y abstención	Contrastar H3 (RF4, RNF2)	Conjunto de validación independiente	Error de calibración esperado (ECE); curva cobertura-precisión
E6. Rostros manipulados	Comprobar la generalidad del diseño	DeepfakeBench (Yan, Zhang, Yuan, et al., 2023)	AUC entre conjuntos de datos

EXPERIMENTO	***PROPÓSITO***	***DATOS***	***MÉTRICAS***
E7. Costo computacional	Valorar la viabilidad del despliegue	Inferencia en GPU y en CPU	Latencia por imagen; memoria

Fuente: Elaboración propia. Como líneas base se usan las ramas individuales, implementadas según Ojha et al. (2023), Tan et al. (2024) y Ricker, Lukovnikov, et al. (2024).

Las particiones se generan con semillas publicadas y ningún modelo ve los generadores de prueba durante el entrenamiento. Para evitar fugas de información, el metaclasificador de M5 se ajusta con un conjunto de validación distinto del que sirvió para entrenar las ramas. Tomando como referencia a Saberi et al. (2024), el sistema se somete también a una purificación por difusión, que funciona como prueba de estrés ante un posible adversario.

DISCUSIÓN

ESPEJO no suma un detector más a los que ya existen, sino que plantea una forma de hacer que esos detectores trabajen juntos. Partimos de la hipótesis de diseño de que los puntos ciegos recogidos en la Tabla 2 casi no coinciden, y que al combinarlos baja la probabilidad de que un generador nuevo logre engañarlos a todos a la vez. La ablación prevista (E4) permitirá ver si esa independencia se cumple o si las ramas tienden a fallar en los mismos casos. Este planteamiento concuerda con lo que Ojha et al. (2023) y Tan et al. (2024) sugieren por caminos diferentes: generaliza mejor aquello que no depende de las huellas de un generador en particular.

La rama de reconstrucción merece un comentario propio, ya que es la que lleva más lejos la idea del título. En ella, el sistema le pregunta al mismo generador si la imagen salió de él (Ricker, Lukovnikov, et al., 2024; Wang et al., 2023). Es una estrategia elegante, aunque también la más vulnerable si cambia la manera de generar imágenes. Justamente por ese riesgo la arquitectura es modular (RNF3). Si una rama queda obsoleta, se puede sustituir sin rehacer todo el sistema, y el protocolo en línea (E2) sirve para identificar el momento en que esto sucede.

Hay dos decisiones que se alejan de lo que suele hacerse. La primera fue emplear un metaclasificador sencillo y no una red profunda de fusión. Se pierde algo de capacidad, pero se gana interpretabilidad y se reduce el riesgo de sobreajuste, lo cual parece prudente si se considera lo heterogéneos que son los datos disponibles (Yan, Zhang, Yuan, et al., 2023). La segunda fue dar al sistema la opción de abstenerse. Si la detección tiene límites teóricos (Saberi et al., 2024), un sistema que siempre emite un dictamen oculta su incertidumbre; en

cambio, uno capaz de decir "no lo sé" la hace visible y deja la decisión en manos de una persona.

Desde el punto de vista ético, el diseño descansa en dos garantías: probabilidades calibradas en vez de etiquetas absolutas y explicaciones visuales que hacen posible revisar cada decisión (Bird y Lotfi, 2024). Con ellas buscamos que el sistema apoye el criterio de periodistas, docentes o peritos, no que lo sustituya. Un falso positivo puede restar valor a una prueba auténtica y un falso negativo puede dar credibilidad a una falsificación. En los dos casos se pone en riesgo la confianza pública en la información, que es justamente el bien que protege el ODS 16.

CONCLUSIÓN

Retomamos la pregunta del inicio. Los 30 estudios revisados indican que es poco probable que un detector aislado logre ser eficaz frente a generadores desconocidos, porque cada principio de detección tiene su propio punto ciego. Como respuesta proponemos ESPEJO, una arquitectura que convierte nueve requisitos en siete módulos. Incluye tres ramas que analizan la imagen desde ángulos diferentes (semántico, estructural y de reconstrucción), un verificador de procedencia, una fusión calibrada con opción de abstención y un módulo que explica cada decisión y funciona como servicio. Dicho de otro modo, la generalización se construye combinando principios y reconociendo abiertamente la incertidumbre.

En el plano práctico, el diseño pone a disposición de universidades, medios y organismos públicos una base abierta y auditable para analizar imágenes dudosas, en línea con el ODS 16. En el plano teórico, transforma en hipótesis contrastables (H1 a H3) una intuición que la literatura ya sugería, pero que no se había probado de manera integrada.

Este estudio tiene limitaciones que es necesario señalar. Hasta que el protocolo se ejecute, las ventajas de la fusión siguen siendo hipótesis. La rama de reconstrucción y el codificador de CLIP aumentan el costo computacional, y eso podría limitar su uso en equipos con pocos recursos (E7). M4 solo resulta útil cuando los desarrolladores incluyen marcas de agua, algo que no ocurre con los modelos abiertos que alguien modifica con fines maliciosos. La arquitectura hereda, por último, los sesgos de los conjuntos de datos públicos, armados casi siempre con contenidos y plataformas de Norteamérica, Europa y Asia oriental.

Como siguientes pasos nos proponemos implementar el prototipo, ejecutar el protocolo y publicar el código y los resultados; incorporar nuevas ramas a medida que aparezcan

generadores con otras arquitecturas; y probar el sistema con imágenes que circulan en plataformas latinoamericanas, donde la evidencia todavía es escasa.

DECLARACIÓN DE USO DE INTELIGENCIA ARTIFICIAL

Durante la preparación de este trabajo usamos herramientas de inteligencia artificial generativa como apoyo en la búsqueda bibliográfica, la estructuración del manuscrito y la redacción. Los autores revisamos, verificamos y editamos todo el contenido, incluidas las referencias, y asumimos plena responsabilidad por su exactitud y originalidad.

CONTRIBUCIÓN DE LOS AUTORES

Danny Fabian Cando Gordon: conceptualización, metodología, diseño de la arquitectura y redacción del borrador original. Guillermo Fernando Argoti Santacruz: revisión de la literatura, validación, y revisión y edición del manuscrito. Marco Yandun: investigación, análisis de los estudios seleccionados, elaboración de tablas y figuras, y revisión crítica del manuscrito. Los tres autores leyeron y aprobaron la versión final.

REFERENCIAS

Bird, J. J., y Lotfi, A. (2024). CIFAKE: Image classification and explainable identification of AI-generated synthetic images. IEEE Access, 12, 15642–15650. https://doi.org/10.1109/ACCESS.2024.3356122

Chen, B., Zeng, J., Yang, J., y Yang, R. (2024). DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. En Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 7621–7639). PMLR. https://proceedings.mlr.press/v235/chen24ay.html

Coccomini, D. A., Messina, N., Gennaro, C., y Falchi, F. (2022). Combining EfficientNet and vision transformers for video deepfake detection. En Image Analysis and Processing – ICIAP 2022 (pp. 219–229). Springer.

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., y Verdoliva, L. (2023). On the detection of synthetic images generated by diffusion models. En IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. https://doi.org/10.1109/ICASSP49357.2023.10095167

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., y Verdoliva, L. (2024). Raising the bar of AI-generated image detection with CLIP. En Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition Workshops (pp. 4356–4366). IEEE. https://doi.org/10.1109/CVPRW63382.2024.00439

Cozzolino, D., Poggi, G., Nießner, M., y Verdoliva, L. (2024). Zero-shot detection of AI-generated images. En Computer Vision – ECCV 2024. Springer.

Epstein, D. C., Jain, I., Wang, O., y Zhang, R. (2023). Online detection of AI-generated images. En Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 382–392). IEEE. https://doi.org/10.1109/ICCVW60793.2023.00045

Fernandez, P., Couairon, G., Jégou, H., Douze, M., y Furon, T. (2023). The stable signature: Rooting watermarks in latent diffusion models. En Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22466–22477). IEEE.

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., Dürmuth, M., y Holz, T. (2024). A representative study on human detection of artificially generated media across countries. En IEEE Symposium on Security and Privacy (pp. 55–73). IEEE.

Haliassos, A., Mira, R., Petridis, S., y Pantic, M. (2022). Leveraging real talking faces via self-supervision for robust forgery detection. En Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14950–14962). IEEE. https://doi.org/10.1109/CVPR52688.2022.01453

Heidari, A., Jafari Navimipour, N., Dag, H., y Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. WIREs Data Mining and Knowledge Discovery, 14(2), Artículo e1520. https://doi.org/10.1002/widm.1520

Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., y Zhao, Y. (2024). Forgery-aware adaptive transformer for generalizable synthetic image detection. En Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10770–10780). IEEE.

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., y Ouyang, W. (2023). Seeing is not always believing: Benchmarking human and model perception of AI-generated images. En Advances in Neural Information Processing Systems (Vol. 36). Curran Associates.

Nightingale, S. J., y Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences, 119(8), Artículo e2120481119. https://doi.org/10.1073/pnas.2120481119

Ojha, U., Li, Y., y Lee, Y. J. (2023). Towards universal fake image detectors that generalize across generative models. En Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 24480–24489). IEEE.

Passos, L. A., Jodas, D., Costa, K. A. P., Souza Júnior, L. A., Rodrigues, D., Del Ser, J., Camacho, D., y Papa, J. P. (2024). A review of deep learning-based approaches for deepfake content detection. Expert Systems, 41(8), Artículo e13570. https://doi.org/10.1111/exsy.13570

Rana, M. S., Nobi, M. N., Murali, B., y Sung, A. H. (2022). Deepfake detection: A systematic literature review. IEEE Access, 10, 25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404

Ricker, J., Damm, S., Holz, T., y Fischer, A. (2024). Towards the detection of diffusion model deepfakes. En Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (Vol. 4, pp. 446–457). SciTePress.

Ricker, J., Lukovnikov, D., y Fischer, A. (2024). AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. En Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9130–9140). IEEE. https://doi.org/10.1109/CVPR52733.2024.00872

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., y Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. En Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684–10695). IEEE.

Saberi, M., Sadasivan, V. S., Rezaei, K., Kumar, A., Chegini, A., Wang, W., y Feizi, S. (2024). Robustness of AI-image detectors: Fundamental limits and practical attacks. En The Twelfth International Conference on Learning Representations (ICLR 2024). https://proceedings.iclr.cc/paper_files/paper/2024/hash/744e09d3213ee53de88f04164a9b06ab-Abstract-Conference.html

Sha, Z., Li, Z., Yu, N., y Zhang, Y. (2023). DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. En Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (pp. 3418–3432). ACM.

Shiohara, K., y Yamasaki, T. (2022). Detecting deepfakes with self-blended images. En Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18720–18729). IEEE.

Tan, C., Zhao, Y., Wei, S., Gu, G., y Wei, Y. (2023). Learning on gradients: Generalized artifacts representation for GAN-generated images detection. En Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12105–12114). IEEE.

Tan, C., Liu, H., Zhao, Y., Wei, S., Gu, G., Liu, P., y Wei, Y. (2024). Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. En Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 28130–28139). IEEE. https://doi.org/10.1109/CVPR52733.2024.02657

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., y Li, H. (2023). DIRE for diffusion-generated image detection. En Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22445–22455). IEEE. https://doi.org/10.1109/ICCV51070.2023.02051

Wen, Y., Kirchenbauer, J., Geiping, J., y Goldstein, T. (2023). Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. En Advances in Neural Information Processing Systems (Vol. 36). Curran Associates.

Yan, Z., Zhang, Y., Fan, Y., y Wu, B. (2023). UCF: Uncovering common features for generalizable deepfake detection. En Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22412–22423). IEEE. https://doi.org/10.1109/ICCV51070.2023.02048

Yan, Z., Zhang, Y., Yuan, X., Lyu, S., y Wu, B. (2023). DeepfakeBench: A comprehensive benchmark of deepfake detection. En Advances in Neural Information Processing Systems (Vol. 36). Curran Associates.

Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., y Wang, Y. (2023). GenImage: A million-scale benchmark for detecting AI-generated image. En Advances in Neural Information Processing Systems (Vol. 36). Curran Associates.

© Los autores. Este artículo se publica en Prisma ODS bajo la Licencia Creative Commons Atribución 4.0 Internacional (CC BY 4.0). Esto permite el uso, distribución y reproducción en cualquier medio, incluidos fines comerciales, siempre que se otorgue la atribución adecuada a los autores y a la fuente original.





: https://doi.org/10.65011/prismaods.v5.i6.440

Cómo citar este artículo (APA 7ª edición):
Cando Gordon, D. F., Argoti Santacruz, G. F., & Yandún Velasteguí, M. A. (2026). Inteligencia Artificial contra Inteligencia Artificial en la Detección de Deepfakes mediante una Arquitectura Híbrida de Redes Neuronales Profundas. *Prisma ODS: Revista Multidisciplinaria Sobre Desarrollo Sostenible*, *5*(6), 465-484. https://doi.org/10.65011/prismaods.v5.i6.440