Pontificia Universidad Javeriana Cali
Facultad de Ingeniería.
Ingeniería de Sistemas y Computación.
Trabajo de Grado.

Comparación del rendimiento de modelos de aprendizaje profundo para distinguir entre imágenes de noticias reales y falsas generadas por IA en redes sociales

Mateo Ramirez Gutierrez

Director: M.Sc. Felipe Palta

22 de Mayo 2026

Resumen

Este trabajo realizó una comparación experimental de seis detectores basados en aprendizaje profundo para distinguir imágenes noticiables reales de imágenes sintéticas generadas por IA, con énfasis en su capacidad de generalización a generadores contemporáneos. Los modelos (UniversalFakeDetect, D^3, DistilDIRE y tres detectores disponibles en Hugging Face) se integraron y adaptaron mediante aprendizaje por transferencia usando el split de entrenamiento de MiRAGeNews, utilizando únicamente la modalidad visual. La comparación se llevó a cabo sobre un conjunto de validación personalizado y balanceado ($n = 100$; 50 reales y 50 sintéticas), y se cuantificó con métricas de clasificación (exactitud, precisión, recall, F1 y ROC-AUC), considerando "fake" como clase positiva, para caracterizar tanto la separabilidad como el perfil de error (falsos positivos vs falsos negativos) relevante en escenarios de verificación periodística. Debido al tamaño limitado y a la selección manual del conjunto externo, estos resultados deben considerarse indicativos de una evidencia inicial pero no equivalen a una validación amplia en escenarios reales de redes sociales. Además, aunque el problema está motivado por el contexto colombiano, la evidencia experimental no se basa en un conjunto de datos específicamente colombiano y por tanto las implicaciones para dicho contexto deben interpretarse con cautela. Los resultados muestran que UniversalFakeDetect y D^3 alcanzaron el mejor desempeño (exactitud 0,91/0,90; ROC-AUC 0,974/0,973), mientras que DistilDIRE presentó un comportamiento cercano al azar (exactitud 0,54; ROC-AUC 0,568) y los modelos de Hugging Face obtuvieron rendimiento intermedio (exactitud 0,68–0,73). En conjunto, estos hallazgos aportan evidencia inicial que puede orientar la selección y calibración de detectores para mitigar la desinformación visual, pero no deben interpretarse como una validación concluyente ni como prueba de impacto directo en contextos específicos. Se reconoce que la evaluación externa presenta limitaciones relevantes: tamaño reducido del conjunto de validación, selección manual de ejemplos y posible sesgo de dominio en las imágenes reales (p. ej., diferencias en fuentes, formatos o contexto). Estudios futuros con muestras más amplias y representativas, evaluación en datos procedentes de redes sociales y análisis del rendimiento por subdominios geográficos son necesarios para establecer recomendaciones operativas.

Palabras Clave: Aprendizaje automático, aprendizaje profundo, inteligencia artificial, detección de imágenes sintéticas, visión por computador, noticias falsas.

Abstract

This study conducted an experimental comparison of six deep learning-based detectors designed to distinguish real newsworthy images from AI-generated synthetic images, with a focus on their ability to generalize to contemporary generators. The models (UniversalFakeDetect, D^3, DistilDIRE, and three detectors available on Hugging Face) were integrated and adapted via transfer learning using the MiRAGeNews training split, utilizing only the visual modality. The comparison was conducted on a customized and balanced validation set ($n = 100$; 50 real and 50 synthetic), and was quantified using classification metrics (accuracy, precision, recall, F1, and ROC-AUC), considering "fake" as the positive class, to characterize both separability and the error profile (false positives vs. false negatives) relevant in journalistic verification scenarios. Due to the limited size and manual selection of the external dataset, these results should be considered indicative of initial evidence but do not amount to a comprehensive validation in real-world social media scenarios. Although the problem is motivated by the Colombian context, the experimental evidence is not based on a specifically Colombian dataset, and therefore the implications for that context should be interpreted with caution. The results show that UniversalFakeDetect and D^3 achieved the best performance (accuracy 0.91/0.90; ROC-AUC 0.974/0.973), while DistilDIRE performed close to random (accuracy 0.54; ROC-AUC 0.568) and the Hugging Face models achieved intermediate performance (accuracy 0.68–0.73). Taken together, these findings provide initial evidence that can guide the selection and calibration of detectors to mitigate visual misinformation, but should not be interpreted as conclusive validation or as proof of direct impact in specific contexts. It is acknowledged that the external evaluation has significant limitations: a small validation set, manual selection of examples, and potential domain bias in the real-world images (e.g., differences in sources, formats, or context). Future studies with larger, more representative samples, evaluation on social media data, and analysis of performance by geographic subdomains are necessary to establish operational recommendations.

Keywords: Machine learning, deep learning, artificial intelligence, AI-generated image detection, computer vision, fake news.

Índice general

Índice de figuras

Índice de cuadros

Introducción

El auge de las inteligencias artificiales generativas ha transformado la manera en que se produce y consume contenido en el entorno digital. La capacidad de generar imágenes y videos con apariencia realista reduce la barrera de entrada para crear piezas visuales persuasivas y de rápida difusión, lo que incrementa el riesgo de que estos recursos se utilicen para apoyar narrativas engañosas o desinformativas en redes sociales, donde además este tipo de contenido tiende a amplificarse con rapidez. Esta problemática afecta directamente la confianza pública en las fuentes de información, amenazando la veracidad y la credibilidad de los medios digitales. En Colombia, donde una proporción significativa de la población participa activamente en redes sociales y donde la capacidad de discernir la autenticidad de contenidos visuales es limitada, el impacto potencial de la desinformación visual adquiere una relevancia particular.

Si bien existen avances importantes en la detección automática de contenido sintético, con métodos que van desde el análisis forense hasta clasificadores de aprendizaje profundo, aún persiste una brecha práctica que dificulta su adopción en escenarios noticiosos reales: los modelos y métricas reportados en la literatura suelen evaluarse bajo conjuntos de datos, generadores y protocolos heterogéneos (con frecuencia centrados en dominios específicos como rostros o configuraciones controladas), lo que limita la comparabilidad directa entre enfoques y hace incierta su generalización a imágenes periodísticas diversas y a generadores de última generación. En consecuencia, ante un caso de uso como la verificación de imágenes noticiables en redes sociales, no es evidente qué arquitecturas ofrecen el mejor compromiso entre desempeño, robustez fuera de dominio y viabilidad operativa bajo restricciones realistas.

Esta brecha motiva la necesidad de una evaluación comparativa: un estudio que contraste, bajo un mismo conjunto de datos noticioso y un protocolo uniforme de refinamiento y validación, el desempeño de modelos de aprendizaje profundo para distinguir entre imágenes reales e imágenes sintéticas. El objetivo central de este trabajo de grado es evaluar comparativamente modelos selectos de clasificación que diferencien imágenes noticiables reales de imágenes noticiables sintéticas, determinando sus alcances y limitaciones según las métricas establecidas para su evaluación. En particular, este trabajo busca responder: ¿Cómo se desempeñan distintos detectores de aprendizaje profundo, refinados bajo un protocolo común, al clasificar imágenes noticiables reales y sintéticas en un conjunto de evaluación pertinente a la actualidad?

Para ello, se seleccionan detectores representativos del estado del arte reciente y modelos disponibles públicamente, y se adaptan al dominio propuesto mediante configuraciones de aprendizaje por transferencia (refinamiento). El procedimiento considera la selección del conjunto de datos, la definición de un protocolo reproducible de entrenamiento/validación y una evaluación consistente sobre un conjunto de validación balanceado, con el fin de obtener evidencia empírica comparable dentro del alcance del protocolo empleado. Los resultados aportan indicios sobre compromisos prácticos entre desempeño, comportamiento fuera del dominio de entrenamiento y viabilidad operativa, sin pretender garantizar generalización absoluta a todas las tecnologías generativas o escenarios posibles. La metodología describe el conjunto de datos, el protocolo de refinamiento y las arquitecturas

evaluadas. Los resultados presentan las métricas y el análisis de patrones de error. Finalmente, en conclusiones se discuten hallazgos, limitaciones y posibles trabajos futuros.

Capítulo 1

Descripción del Problema

1.1. Planteamiento del Problema

La adopción masiva de modelos de inteligencia artificial (IA) generativa desde 2023 ha incrementado la disponibilidad de contenido audiovisual sintético, en particular imágenes y videos de alta verosimilitud, susceptible de ser publicado y amplificado en redes sociales. Este panorama amplifica los riesgos asociados a la desinformación visual: la circulación de piezas gráficas que, al presentarse como evidencia de un hecho noticioso, pueden influir en la percepción pública y dificultar el acceso oportuno a información verificada.

En plataformas donde la información se consume y comparte a gran velocidad, la visibilidad diferencial de este tipo de contenidos constituye un factor crítico. En X (anteriormente Twitter), publicaciones asociadas a desinformación generada por IA han mostrado mayor propagación que publicaciones legítimas, con 10.81 % más retweets, 34.16 % más *likes* y 10.32 % más impresiones [13]. En consecuencia, aun cuando existan fuentes confiables, la exposición cotidiana del público se ve mediada por dinámicas algorítmicas y sociales que priorizan contenidos altamente atractivos, no necesariamente auténticos.

Particularmente en Colombia, para 2025 se estiman 36.8 millones de usuarios activos de redes sociales, equivalentes al 69.2 % de la población total [14]. Sin embargo, la participación activa en estas plataformas no implica una alta capacidad de discernimiento: en un estudio de 2024, los colombianos diferenciaron correctamente la autenticidad de contenidos visuales solo el 56.9 % de las veces [15]. Esta brecha, combinada con la alta visibilidad del contenido sintético, incrementa la probabilidad de que imágenes no auténticas sean consumidas y compartidas sin cuestionar su autenticidad.

Para enfrentar esta amenaza, la literatura ha propuesto métodos de detección que abarcan desde el análisis forense hasta clasificadores de aprendizaje automático y marcos multimodales. En particular, los enfoques de aprendizaje profundo suelen explotar huellas forenses del dominio visual, tales como inconsistencias a nivel de píxeles, artefactos de color o anomalías de coocurrencia, y entrenar clasificadores supervisados, principalmente redes neuronales convolucionales (CNN) y transformadores de visión (ViT) [3]. No obstante, la evidencia disponible presenta una brecha metodológica: los resultados se reportan sobre benchmarks y protocolos heterogéneos, con particiones, preprocesamientos y métricas no equivalentes, y con frecuencia en dominios que no representan el contexto noticioso contemporáneo. Esto dificulta establecer comparaciones controladas y, por tanto, identificar qué detectores ofrecen mayor robustez y generalización en escenarios pertinentes a la actualidad.

En consecuencia, la detección consistente de imágenes sintéticas en el corto y mediano plazo sigue siendo incierta, especialmente ante la evolución continua de los modelos generativos. Por ello, se requiere una comparación sistemática de detectores de aprendizaje profundo bajo un protocolo común y sobre un dominio noticioso, que permita estimar de manera justa su rendimiento y caracterizar patrones de error. Mediante este trabajo de grado, se propone realizar un benchmarking comparativo de modelos de aprendizaje profundo representativos del estado del arte, buscando el mejor desempeño en la clasificación de imágenes noticiables reales y sintéticas para sentar bases metodológicas que faciliten el desarrollo de aplicaciones orientadas al contexto colombiano.

1.1.1. Formulación

¿Cómo se desempeñan distintos detectores de aprendizaje profundo, refinados bajo un protocolo común, al clasificar imágenes noticiables reales y sintéticas en un conjunto de evaluación pertinente a la actualidad?

1.1.2. Sistematización

Para resolver esta pregunta, se deben tener en cuenta algunas subpreguntas:

- ¿Qué criterios de selección justifican la elección de los modelos con los que se pueda realizar una comparativa en la clasificación de imágenes noticiables reales e imágenes noticiables sintéticas?
- ¿Cómo seleccionar un conjunto de datos de entrenamiento adecuado, sobre el cuál se puedan refinar los modelos para realizar tareas de clasificación sobre imágenes?
- ¿Qué protocolo establecer para el refinamiento y validación de los modelos de aprendizaje profundo previamente elegidos al contexto del conjunto de datos?
- ¿Cómo crear un conjunto de validación que permita evaluar la generalización de los modelos a imágenes noticiables reales y sintéticas con características relevantes al dominio periodístico actual?
- ¿Cómo varía el rendimiento de los modelos entre evaluaciones y qué patrones de error evidencian sus fortalezas y debilidades?

1.2. Objetivos

1.2.1. Objetivo General

Evaluar comparativamente modelos de clasificación selectos que diferencien imágenes noticiables reales de imágenes noticiables sintéticas, utilizando técnicas de aprendizaje profundo y determinando sus alcances y limitaciones según las métricas establecidas para su evaluación.

1.2.2. Objetivos Específicos

1. Identificar los modelos de aprendizaje de máquina profundo más relevantes en la literatura científica, según criterios de selección definidos, que se enfoquen en la clasificación de imágenes noticiables reales e imágenes noticiables sintéticas.

2. Determinar un conjunto de datos pertinente sobre imágenes noticiables reales e imágenes noticiables sintéticas para realizar potenciales entrenamientos de clasificación.

3. Adaptar al menos tres (3) modelos de clasificación de imágenes noticiables reales e imágenes noticiables sintéticas sobre el conjunto de datos determinado.

4. Diseñar un conjunto de validación personalizado para validar la capacidad de generalización de los modelos a imágenes noticiables reales y sintéticas.

5. Analizar los alcances y limitaciones de los modelos de clasificación de imágenes entrenados para distinguir entre imágenes noticiables reales e imágenes noticiables sintéticas según su rendimiento.

1.3. Justificación

A nivel social, la proliferación de imágenes sintéticas de alta verosimilitud amplifica la desinformación visual en plataformas donde el contenido se consume y comparte con rapidez. En este escenario, las imágenes noticiables son especialmente sensibles: suelen circular como "evidencia" de hechos, por lo que errores de clasificación pueden tener consecuencias directas sobre la confianza pública y la toma de decisiones informadas.

Desde el ámbito técnico, La literatura reciente sobre detección de imágenes generadas por IA reporta avances importantes, pero también limitaciones persistentes que impiden una adopción robusta en escenarios reales. En particular, varios estudios se apoyan en conjuntos de datos centrados en rostros o escenarios controlados, donde se alcanzan precisiones superiores al 90 %, pero el rendimiento se degrada cuando se evalúa fuera del dominio de entrenamiento o frente a manipulaciones no vistas [16, 17]. Esto es relevante para el dominio noticioso, ya que en prensa y redes sociales las imágenes tienden a ser heterogéneas (lugares, objetos, eventos, escenas) y varían con rapidez en estilo y procedencia.

En términos de aporte metodológico a la literatura existente, cabe resaltar que incluso los trabajos orientados a redes sociales reconocen que los recursos disponibles aún no capturan plenamente la diversidad, el realismo y la dinámica del ecosistema de plataformas. Benchmarks recientes como So-Fake-Set [18] y SID-Set [19] aumentan escala y realismo, pero subrayan dificultades para generalizar a tecnologías generativas no vistas y a escenarios fuera de laboratorio. Este punto es crítico: a medida que aparecen generadores nuevos, los detectores pueden volverse frágiles si aprenden huellas específicas de un conjunto de modelos generativos o de un preprocesamiento particular; por tanto, se requiere una evaluación externa bajo un protocolo común que estime de forma más justa la capacidad de generalización.

En este marco, la evolución de la calidad visual de los modelos generativos refuerza la necesidad de evaluación, pero debe interpretarse con cautela. El puntaje FID (menor es mejor) aproxima similitud estadística con fotografías reales, pero no mide directamente si una imagen engañará a humanos ni garantiza, por sí solo, que un detector fallará. Aun así, la disminución reportada (por ejemplo, de 22.4 en 2021 a 2.6 en 2024 para modelos líderes) [20] sugiere un incremento en realismo que puede reducir la separabilidad entre clases y exigir detectores con mejor generalización.

En consecuencia, el aporte de este proyecto no se limita a constatar la existencia del problema, sino a ofrecer una comparación controlada y actualizada de detectores de aprendizaje profundo sobre imágenes noticiables reales y sintéticas generadas por IA, usando un mismo conjunto de datos noticioso y un protocolo uniforme de entrenamiento, validación y prueba. Con ello, el trabajo cubre el vacío específico de evaluar externamente, en un dominio noticioso pertinente y con imágenes sintéticas recientes, qué tan bien generalizan detectores refinados frente a variaciones fuera del dominio y generadores no vistos.

1.4. Delimitaciones y Alcances

- El problema de investigación de este trabajo distingue entre (i) *imagen sintética*, entendida como una imagen generada por un modelo de IA (total o parcialmente), (ii) *imagen falsa*, como una imagen presentada de manera engañosa en un contexto noticioso (lo cual puede involucrar generación sintética, edición o recontextualización), (iii) *noticia falsa*, que alude a la veracidad del relato completo y (iv) *desinformación visual*, como el uso de piezas visuales para inducir interpretaciones erróneas. En este sentido, el objeto de estudio no es verificar si una noticia completa es verdadera o falsa, sino clasificar la autenticidad visual de imágenes noticiables (reales vs. sintéticas) que acompañan narrativas informativas.

- El alcance del proyecto se centra en detectores visuales; aunque MiRAGeNews es multimodal, en este trabajo se utilizan únicamente las imágenes y la etiqueta binaria asociada al par imagen-subtítulo como extensión de autenticidad visual. Por ende, el análisis multimodal (consistencia imagen-texto), verificación semántica de la narrativa, uso de metadatos, procedencia, rastreo de fuentes, detección en tiempo real (latencia a nivel de despliegue) y evaluación con usuarios no se incluyen para efectos de este estudio.

- Se evalúan seis (6) modelos de detección: UniversalFakeDetect, D^3, DistilDIRE, Smogy/SMOGY-Ai-images-detector, Ateeqq/ai-vs-human-image-detector y prithivMLmods/deepfake-detector-model-v1.

- Para adaptar los modelos al dominio noticioso, el refinamiento se realiza con el split *train* de MiRAGeNews [11], utilizando 10.000 imágenes (balanceadas entre real y sintética) y trabajando únicamente con la modalidad visual. Los splits *validation* y varios *test* de MiRAGeNews no se usan como evaluación principal del trabajo, se consultan sólo como referencia y análisis exploratorio, ya que pueden presentar debilidades metodológicas y su contexto temporal

(imágenes y generadores de 2023) no representa de forma suficiente el ecosistema de generación sintética más reciente.

- La comparación reportada de los seis detectores se realiza sobre un conjunto de validación personalizado y balanceado, con un tamaño de 100 imágenes (50 reales y 50 sintéticas). Este conjunto es externo al refinamiento: no se utiliza para entrenar, sino para estimar generalización bajo condiciones más cercanas al escenario objetivo, incorporando imágenes reales de dominio noticioso y un split sintético generado con modelos contemporáneos. Dado su tamaño y curaduría, su carácter es exploratorio y sus métricas deben interpretarse como evidencia inicial.
- En el repositorio público y metadatos del dataset no se identificó una licencia explícita y unificada para MiRAGeNews. En consecuencia, el uso del conjunto de datos por fuera de fines académicos se asume como un riesgo y se condiciona estrictamente el cumplimiento de (i) la citación explícita del recurso, (ii) la no redistribución del dataset ni publicación de derivados que lo reempaqueten, y (iii) el uso limitado a entrenamiento y evaluación internos dentro del proyecto, respetando los términos aplicables de las fuentes subyacentes.

Capítulo 2

Marco Teórico

2.1. Conceptos básicos

Este trabajo se centra en la detección automática de imágenes noticiables generadas por inteligencia artificial, formulada como un problema de clasificación binaria entre imágenes reales e imágenes de la clase fake (imágenes sintéticas). En consecuencia, este marco teórico se organiza alrededor de decisiones metodológicas del proyecto: (i) delimitar qué se entiende por contenido visual sintético en el contexto de desinformación, (ii) justificar el uso de visión por computadora para operacionalizar la autenticidad como una tarea de clasificación, y (iii) fundamentar la selección de modelos, estrategias de entrenamiento y métricas con las que se compararán los enfoques.

2.1.1. Deepfakes

Un deepfake se define como un tipo de medio sintético en el que contenidos audiovisuales —habitualmente imágenes, vídeo o audio de personas reales o verosímiles— son generados o manipulados mediante técnicas avanzadas de inteligencia artificial, en particular aprendizaje profundo, con el fin de producir representaciones altamente realistas pero falsas que pueden llegar a confundirse con material genuino [21, 22].

El término abarca tanto la manipulación de material preexistente como la generación completamente sintética de contenido. En términos técnicos, estas transformaciones se implementan mediante modelos generativos entrenados para aproximar la distribución de los datos reales, de modo que las salidas resulten verosímiles y difíciles de distinguir sin asistencia computacional [23]. Aunque el fenómeno suele asociarse a vídeo y audio, en este proyecto se restringe el análisis al caso de imágenes, ya que el objetivo experimental es comparar modelos de clasificación visual para distinguir imágenes reales frente a imágenes fake.

2.1.2. Contenido visual generado por IA

El contenido visual generado por IA comprende imágenes producidas, manipuladas o modificadas por algoritmos de inteligencia artificial (en lugar de autores humanos directos), típicamente mediante modelos de aprendizaje automático entrenados con grandes conjuntos de datos [24]. En este trabajo, este concepto se operacionaliza como la clase fake del conjunto de datos utilizado: imágenes sintéticas que circulan en contextos noticiosos y que deben distinguirse de imágenes reales mediante clasificadores visuales.

2.2. Visión por computadora

La visión por computadora es el campo interdisciplinar de la informática y la inteligencia artificial que estudia cómo dotar a las máquinas de la capacidad de adquirir, procesar y comprender información visual de manera análoga, aunque no idéntica, a la percepción humana. En esta disciplina se desarrollan modelos teóricos, representaciones matemáticas y algoritmos que permiten transformar datos de imagen en descripciones numéricas o simbólicas del mundo, de forma que un sistema artificial pueda identificar objetos, escenas, acciones y relaciones espaciales, tomar decisiones o realizar tareas que, en los humanos, dependen del sistema visual [25].

Desde la ingeniería, la visión por computadora se concibe como el intento de automatizar tareas que tradicionalmente requerían la interpretación visual humana, como la inspección industrial, el reconocimiento de rostros, la conducción autónoma o el análisis médico de imágenes [26, 27]. En términos de flujo de procesamiento, un sistema de visión por computadora adquiere imágenes, aplica preprocesamiento, extrae características, realiza análisis y razonamiento, y finalmente produce una salida interpretable (por ejemplo, una clase, una localización o una segmentación) [27].

En el ámbito de la desinformación visual, el cual va desde imágenes cuyo contexto ha sido manipulado o eliminado hasta vídeos sintéticamente generados, la visión por computadora se ha convertido en una herramienta central tanto para la detección automática del engaño como para el análisis forense de su origen. De esta forma, las labores se centran en la detección de contenido sintético empleando modelos de aprendizaje profundo capaces de identificar artefactos sutiles introducidos por los procesos de generación o edición, tales como inconsistencias de textura, anomalías de iluminación o patrones espectrales atípicos en el dominio de la frecuencia [28].

En este proyecto, la visión por computadora constituye el marco adecuado porque el problema se formula directamente sobre evidencia visual: distinguir si una imagen noticiable corresponde a una captura real del mundo o a una síntesis generada por IA. Dado que este tipo de contenido circula masivamente en redes sociales, puede alcanzar alta visibilidad y viralidad [13], y al mismo tiempo puede superar la capacidad humana de discernimiento de autenticidad en una proporción relevante de casos [15]. En consecuencia, se requieren mecanismos automáticos, escalables y reproducibles para analizar grandes volúmenes de imágenes y producir una decisión de autenticidad de forma consistente.

Metodológicamente, este proyecto aterriza la visión por computadora a una tarea de clasificación de imágenes en un escenario binario (real vs. fake), apoyándose en modelos que aprenden representaciones discriminativas a partir de datos. Este planteamiento permite comparar enfoques del estado del arte y evaluar sus fortalezas y limitaciones bajo un mismo protocolo experimental, alineado con el objetivo de benchmarking propuesto [3].

2.2.1. Clasificación de imágenes

La clasificación de imágenes, como subcampo de la visión por computadora, puede definirse de manera académica como el proceso mediante el cual un sistema computacional analiza el contenido visual de una imagen y le asigna una o varias etiquetas pertenecientes a un conjunto finito y predefinido de categorías semánticas, en función de los patrones que reconoce en los datos de

píxeles [29, 30].

En este contexto, la imagen se considera como una realización de una variable aleatoria de alta dimensión, y la tarea de clasificación se formula como un problema de aprendizaje supervisado en el que, a partir de un conjunto de ejemplos de entrenamiento anotados, se estima una función de decisión que mapea cada imagen x a una clase y perteneciente a un espacio discreto de etiquetas [30].

En el problema de la desinformación visual, la clasificación de imágenes se utiliza para distinguir entre contenido auténtico y contenido manipulado o sintético. Esta distinción puede formularse con múltiples categorías (por ejemplo, "imagen manipulada" o "imagen fuera de contexto") o con un esquema binario "real" frente a "sintética" [31]. En este proyecto se adopta el esquema binario, alineado con la etiqueta fake del conjunto de datos, para facilitar una comparación homogénea entre arquitecturas bajo un mismo protocolo experimental.

2.3. Teoría de aprendizaje automático

El aprendizaje automático es un subcampo de la inteligencia artificial (IA) que busca aprender funciones de decisión a partir de datos para predecir o clasificar sin reglas explícitas diseñadas manualmente [32]. En este proyecto, la tarea de interés se modela como un problema de clasificación supervisada: a partir de ejemplos etiquetados de imágenes noticiables reales e imágenes noticiables fake, se entrena y compara un conjunto de modelos con el objetivo de generalizar a nuevas imágenes. Esta formulación es adecuada porque traduce la autenticidad visual a una decisión operacional medible, habilita una evaluación comparativa controlada entre arquitecturas bajo métricas comunes, y permite refinar modelos preexistentes sobre un conjunto de datos específico, en lugar de proponer un detector ad-hoc difícil de contrastar.

2.3.1. Aprendizaje supervisado

El aprendizaje supervisado es un paradigma de aprendizaje automático en el que se entrena un modelo con conjuntos de datos etiquetados, es decir, pares de entrada–salida conocidos, con el objetivo de que el sistema aprenda una función que relacione ambas y pueda predecir correctamente la salida para nuevas entradas no vistas. En el contexto de clasificación de imágenes, este paradigma se materializa típicamente mediante redes neuronales profundas convolucionales (CNN) o transformadores visuales (ViT), que ajustan sus parámetros internos minimizando una función de pérdida que cuantifica el error entre las etiquetas verdaderas y las predicciones del modelo [33, 34].

Este enfoque resulta pertinente para la problemática descrita porque la salida de interés (autenticidad visual) puede definirse como una etiqueta discreta y verificable en el marco del conjunto de datos utilizado (real vs. fake). Al entrenar con etiquetas, el modelo aprende una frontera de decisión directamente optimizada para la distinción que se desea evaluar, y su desempeño puede contrastarse de manera objetiva mediante métricas como exactitud, precisión, recall, F1, AUC-ROC y matriz de confusión. Además, al mantener el mismo conjunto de datos y protocolo de entrenamiento/validación, el aprendizaje supervisado facilita una comparación justa entre arquitecturas

seleccionadas.

2.3.2. Redes neuronales

Las redes neuronales son modelos computacionales de aprendizaje automático inspirados en el funcionamiento del cerebro humano, compuestos por nodos o "neuronas" artificiales interconectadas que procesan información y ajustan sus conexiones a partir de datos para reconocer patrones, estimar relaciones y producir una salida [35]. Pueden entenderse como sistemas de procesamiento y aprendizaje basados en capas de unidades simples, conectadas entre sí mediante pesos que se modifican durante el entrenamiento, lo que les permite aproximar funciones complejas y resolver tareas como clasificación, predicción y reconocimiento de patrones. La variación en el tamaño de una red neuronal se puede visualizar en la Figura 2.1.

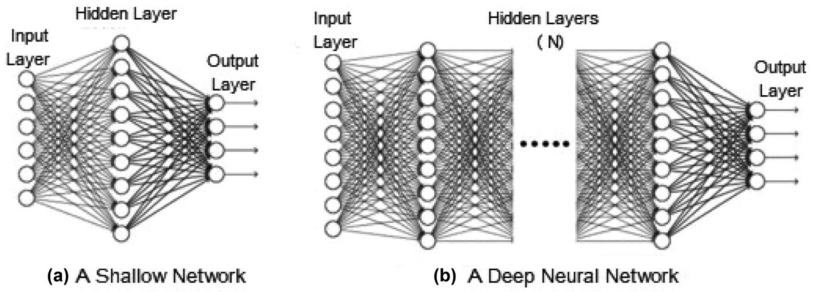


Figura 2.1: Comparativa entre redes neuronales según su tamaño (Fuente: [4]).

En este trabajo, las redes neuronales se concretan en arquitecturas profundas (CNN y ViT) utilizadas como extractores de representaciones para la clasificación de imágenes reales frente a fake.

2.3.3. Aprendizaje profundo

El aprendizaje profundo se define como una rama del aprendizaje automático basada en redes neuronales profundas que aprende representaciones jerárquicas a partir de grandes volúmenes de datos, de modo que el sistema puede identificar patrones complejos y realizar tareas como clasificación, predicción o generación de contenido con alto grado de precisión [36]. En el ámbito de la desinformación visual, este enfoque ha sido especialmente importante porque permite modelar rasgos sutiles que los métodos manuales o superficiales suelen pasar por alto, como inconsistencias en bordes faciales, texturas y artefactos de síntesis que aparecen en imágenes manipuladas o generadas.

Para la distinción entre imágenes noticiables reales y sintéticas, el aprendizaje profundo es adecuado por dos razones principales. Primero, evita depender exclusivamente de reglas o descriptores forenses diseñados manualmente, que pueden quedarse cortos cuando los modelos generativos cambian sus patrones de error; en su lugar, aprende características discriminativas directamente de los

datos. Segundo, los modelos profundos contemporáneos usados en esta línea (CNN y ViT) constituyen el estado del arte en clasificación visual y han sido empleados ampliamente para detección de contenido sintético [3]. Esto es coherente con el objetivo del proyecto de comparar y evaluar modelos relevantes reportados en la literatura.

2.3.4. Arquitecturas en el aprendizaje profundo

Una arquitectura de un modelo de aprendizaje profundo contiene el diseño que determina cómo el sistema extrae y combina representaciones de diferentes niveles de abstracción a medida que la información avanza por sus capas, transformando datos de entrada en una salida útil, como una clasificación o una predicción [37]. Para efectos de la transformación de datos de tipo imagen, hay dos arquitecturas que se destacan por servir como base para modelos más detallados: las redes neuronales convolucionales y los transformadores visuales.

2.3.4.1. Redes neuronales convolucionales (CNN)

Las redes neuronales convolucionales (CNN) son arquitecturas profundas especializadas en datos con estructura espacial, sobre todo imágenes. En otros términos, una CNN es una red neuronal que emplea operaciones de convolución para detectar patrones locales como bordes, texturas, formas y composiciones visuales, combinando luego esas señales en representaciones cada vez más complejas hasta llegar a una decisión final [37]. Su lógica de operación se apoya en filtros que recorren la imagen y capturan rasgos visuales relevantes; después, capas de agrupación o reducción de dimensiones ayudan a conservar lo más importante y a hacer el modelo más eficiente. Una visualización de esta arquitectura se puede ver en la Figura 2.2.

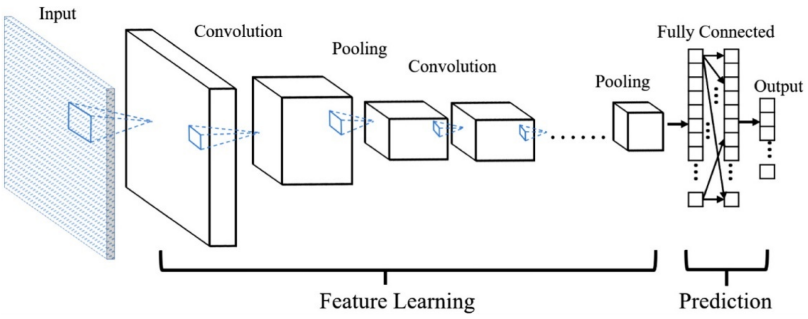


Figura 2.2: Ejemplo de una arquitectura CNN (Fuente: [5]).

En el contexto de imágenes empleadas con fines de desinformación, los modelos basados en la arquitectura CNN resultan útiles porque pueden aprender huellas visuales de manipulación que no siempre son obvias para una persona, como inconsistencias en el contorno de un rostro, artefactos de síntesis, texturas anómalas o patrones de compresión y edición asociados con falsificación digital. Debido a esto, se han usado CNNs en gran parte de la literatura temprana y reciente en detección

de contenido visual manipulado y/o sintético como base para clasificar imágenes reales frente a alteradas.

2.3.4.2. Transformadores visuales (ViT)

Los transformadores visuales (ViT) son una arquitectura que adapta la familia de transformadores al análisis de imágenes. En lugar de extraer rasgos visuales mediante convoluciones locales, un ViT divide la imagen en parches, los convierte en secuencias comparables a "tokens" y aprende relaciones entre ellos mediante mecanismos de atención, de modo que el modelo puede considerar dependencias globales entre regiones distantes de la imagen desde las primeras etapas del procesamiento [37]. Un ejemplo visual de esta arquitectura se puede ver en la Figura 2.3.

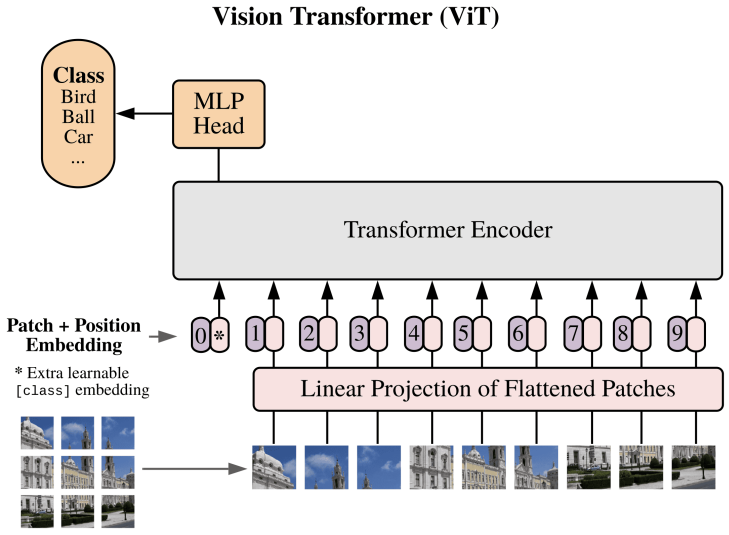


Figura 2.3: Ejemplo de una arquitectura ViT (Fuente: [6]).

En la práctica, esto significa que un ViT no se centra solo en detectar detalles locales, sino también en cómo distintas partes de la imagen se relacionan entre sí para formar una representación coherente o incoherente del contenido visual. En imágenes empleadas con fines de desinformación, esa capacidad puede ser muy valiosa porque ciertas manipulaciones afectan la consistencia global de la escena, la armonía entre fondo y sujeto, o la relación espacial entre elementos que, vistos en conjunto, delatan una edición o una generación sintética.

2.3.5. Aprendizaje por transferencia

El aprendizaje por transferencia es una técnica de aprendizaje automático en la que se aprovecha el conocimiento adquirido al entrenar un modelo en una tarea o conjunto de datos para mejorar el aprendizaje o el rendimiento en una tarea o conjunto de datos diferente, pero relacionado.

En términos generales, la transferencia se implementa iniciando el entrenamiento desde un modelo preentrenado en un dominio de origen (normalmente un conjunto de datos grande y diverso), de manera que el modelo ya disponga de representaciones visuales útiles en niveles bajos y medios (bordes, texturas, patrones y composiciones). Posteriormente, ese conocimiento se adapta al dominio objetivo ajustando (fine-tuning) parte o la totalidad de sus parámetros en el nuevo conjunto de datos, en lugar de entrenar desde cero [38].

En este trabajo, el uso de aprendizaje por transferencia es metodológicamente razonable por varias razones. Primero, el conjunto de datos objetivo disponible para el refinamiento es de tamaño acotado en comparación con los repositorios masivos usados para entrenar modelos fundacionales; por ejemplo, se plantea trabajar con un subconjunto de 10.000 imágenes de entrenamiento [11]. En este escenario, entrenar modelos profundos desde cero incrementa el riesgo de sobreajuste y exige un costo computacional mayor (más épocas y más datos para estabilizar el aprendizaje). La transferencia permite acelerar la convergencia y mejorar la generalización al partir de pesos ya informativos.

Segundo, la transferencia facilita la adaptación al contexto noticioso: aunque los preentrenamientos generales capturan patrones visuales amplios, el dominio objetivo puede incluir distribuciones particulares de imágenes (composición, texto incrustado, estilos de captura, compresión y recortes típicos de redes sociales). La técnica permite ajustar las capas finales y/o la atención del modelo para que la separación real vs. sintética refleje mejor ese dominio.

Tercero, desde la perspectiva de benchmarking, emplear aprendizaje por transferencia contribuye a la comparabilidad entre modelos. Al utilizar puntos de partida estándar (pesos preentrenados) y un protocolo homogéneo de refinamiento (misma división de datos, mismas métricas y criterios de parada), se reduce la variabilidad debida a inicializaciones aleatorias y se enfoca la comparación en las diferencias propias de cada arquitectura. Esto es especialmente importante cuando el objetivo no es maximizar un único modelo, sino contrastar el desempeño de varios enfoques bajo condiciones controladas.

2.3.6. Métricas para evaluar el desempeño

En este proyecto se adopta la convención de tomar como clase positiva la etiqueta fake (imagen sintética). Esta decisión fija el significado de TP/FP/TN/FN y, por tanto, la interpretación de métricas como precisión y recall en todo el documento.

- **Matriz de confusión**

  La matriz de confusión contrasta el número de instancias de cada clase real frente a las clases predichas por el modelo, permitiendo visualizar los aciertos (verdaderos positivos, TP; verdaderos negativos, TN) y los errores (falsos positivos, FP; falsos negativos, FN) [39]. Un ejemplo se puede ver en la Figura 2.4.

  Para esta comparativa, la matriz de confusión es clave porque hace explícito qué se está equivocando el modelo. Por ejemplo, si se toma como clase positiva fake, un FN implica que una imagen fake fue clasificada como real, lo cual es especialmente problemático porque deja

pasar contenido potencialmente engañoso. En cambio, un FP implica marcar como fake una imagen real, lo cual puede erosionar la confianza en el sistema, inducir a descartar contenido legítimo o generar sobrecarga al requerir verificación manual. Este análisis por tipo de error permite justificar fortalezas y debilidades más allá de una sola cifra global.

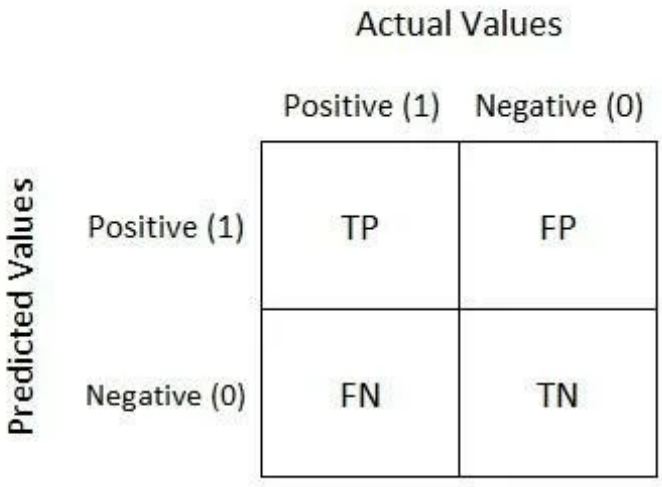


Figura 2.4: Tabla de matriz de confusión.

- **Pérdida (Loss)**

  La pérdida es una función matemática que cuantifica el desacuerdo entre las predicciones del modelo y los valores verdaderos, asignando un costo a cada error para guiar el proceso de optimización durante el entrenamiento [40]. Esta métrica es fundamental porque el entrenamiento busca minimizarla; sin embargo, una pérdida baja no garantiza por sí sola el mejor desempeño en términos de interpretación (por ejemplo, puede existir sobreajuste).

  En el contexto de este trabajo, seguir la evolución de la pérdida en entrenamiento y validación ayuda a detectar convergencia y estabilidad del entrenamiento, identificar sobreajuste cuando la pérdida de entrenamiento disminuye pero la de validación empeora, y comparar qué tan "difícil" resulta para cada modelo aprender a separar imágenes reales de sintéticas en el conjunto de datos seleccionado.

  En modelos de clasificación como los empleados en este proyecto, la función de pérdida más común es la entropía cruzada, al medir el error entre probabilidades predichas y etiquetas reales [40]. En este caso, una pérdida de entropía cruzada baja indica que el modelo está asignando altas probabilidades a las clases correctas, lo cual es deseable para una clasificación precisa de imágenes reales y sintéticas.

- **Exactitud (Accuracy)**

La exactitud es la proporción de predicciones correctas respecto del total de observaciones [39].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Para el benchmarking, la exactitud ofrece una lectura rápida del rendimiento global y facilita comparar modelos bajo el mismo protocolo experimental. No obstante, es importante interpretarla con cautela si existe desbalance de clases (por ejemplo, si hay muchas más imágenes reales que fake, o viceversa), ya que un modelo podría obtener una exactitud alta simplemente favoreciendo la clase mayoritaria.

- **Precisión (Precision)**

  La precisión es la fracción de predicciones positivas que son realmente positivas, midiendo la fiabilidad de las etiquetas positivas emitidas por el modelo [39].

  $$Precision = \frac{TP}{TP + FP}$$

  En este proyecto, si se define como positiva la clase fake, una alta precisión implica que cuando el modelo alerta que una imagen es fake, usualmente acierta. Esto es relevante porque reduce falsas alarmas (FP), evita desacreditar contenido real y mejora la aceptación de la herramienta por parte de usuarios o analistas. En un escenario de verificación, también significa que el esfuerzo humano se invierte en casos con mayor probabilidad de ser realmente fake.

- **Exhaustividad (Recall)**

  El recall (o sensibilidad) es la proporción de instancias realmente positivas que el modelo identifica correctamente como tales [39].

  $$Recall = \frac{TP}{TP + FN}$$

  Para la detección de imágenes fake, el recall es crítico porque mide cuántas imágenes fake se logran detectar. Un recall bajo implica muchos FN (fake clasificadas como reales), lo cual es el caso más riesgoso para la problemática de desinformación: el contenido potencialmente engañoso pasaría el filtro. Por esto, al comparar modelos, el recall ayuda a priorizar aquellos que "dejan pasar" menos fake.

- **F1-score**

  El F1-score combina precisión y recall mediante su media armónica [39].

  $$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Esta métrica es valiosa en el proyecto porque resume el compromiso entre FP y FN en un solo número, penalizando situaciones donde una de las dos métricas sea muy baja. En tareas donde tanto evitar falsas alarmas como capturar la mayor cantidad de fake importa, el F1 facilita una comparación más equilibrada entre modelos, especialmente si hay desbalance entre clases.

- **Curva Precisión–Recall (PR) y AUC-PR**

  La curva precisión–recall (PR) grafica la precisión frente al recall para todos los posibles umbrales de decisión. Esta curva es especialmente informativa cuando la prevalencia de la clase positiva es baja o existe desbalance entre clases, ya que permite observar cómo aumenta el número de falsos positivos al buscar mayor cobertura (recall) de fake [39]. De forma análoga a AUC-ROC, el *área bajo la curva PR* (AUC-PR) o la precisión promedio resume el desempeño a través de umbrales, pero con un enfoque más sensible a la clase positiva en escenarios desbalanceados.

- **Curva ROC (Receiver Operating Characteristic)**

  La curva ROC traza la tasa de verdaderos positivos (TPR) contra la tasa de falsos positivos (FPR) para todos los posibles umbrales de clasificación, ilustrando la compensación entre detectar correctamente los positivos y evitar clasificar incorrectamente los negativos [41]. En este trabajo, la ROC es útil porque muchos modelos producen una puntuación o probabilidad (no solo una etiqueta). Cambiar el umbral permite ajustar el comportamiento del sistema según el objetivo: por ejemplo, exigir mayor recall (ser más "estricto" marcando fake) o mayor precisión (ser más "conservador" para no marcar reales). La curva hace visible este intercambio y permite seleccionar umbrales coherentes con el costo relativo de FP y FN en la detección de desinformación visual.

- **Área Bajo la Curva ROC (AUC-ROC)**

  El AUC-ROC es una métrica escalar que cuantifica la capacidad discriminante global del modelo, con valores entre 0,5 (sin capacidad predictiva) y 1,0 (capacidad perfecta) [41]. Para el benchmarking, el AUC-ROC ofrece una comparación agregada entre modelos al resumir su desempeño a través de todos los umbrales posibles. Esto es útil cuando se desea evaluar la "separabilidad" intrínseca entre imágenes reales y fake, aun antes de fijar un umbral operativo. No obstante, un AUC alto no implica necesariamente un buen desempeño en un umbral concreto: por ejemplo, se puede observar AUC elevado y recall bajo si el umbral operativo se fija de forma demasiado conservadora. Por ello, en este trabajo se recomienda interpretar AUC-ROC junto con la matriz de confusión, métricas puntuales (precisión/recall/F1) y una discusión explícita del criterio de selección o calibración del umbral.

- **Conteos e incertidumbre estadística**

  Finalmente, para que la comparación sea interpretable, conviene reportar los conteos (TP, FP, TN, FN) junto con las métricas agregadas. Esto es especialmente importante cuando el conjunto de prueba es pequeño (por ejemplo, 100 imágenes), ya que variaciones de pocos casos

pueden cambiar de manera apreciable valores como precisión, recall o F1. En estas condiciones, se recomienda acompañar las métricas con estimaciones de incertidumbre (por ejemplo, intervalos de confianza para proporciones y procedimientos de remuestreo para AUC), de modo que las conclusiones reflejen la variabilidad esperable y no solo un punto estimado.

2.4. Trabajos Relacionados

Los trabajos relacionados se organizan aquí con el objetivo de justificar vacíos concretos que este proyecto aborda parcialmente. En particular, se consideran: (i) detectores supervisados que tratan la autenticidad como clasificación binaria real vs. sintética y emplean aprendizaje por transferencia; (ii) propuestas que discuten el compromiso entre rendimiento y eficiencia mediante arquitecturas ligeras; y (iii) revisiones que sistematizan tendencias (GAN vs. difusión; dominio espacial vs. frecuencia) y resaltan problemas persistentes como la generalización a generadores no vistos y el sesgo de dominio. Estos ejes convergen en la motivación central del proyecto: evaluar y comparar modelos en un dominio noticioso (redes sociales) bajo un protocolo homogéneo.

2.4.1. AI vs. AI: Can AI Detect AI-Generated Images? [1]

El trabajo de Baraheem y Nguyen propone un marco de detección de imágenes sintéticas basado en redes neuronales convolucionales (CNN) que distingue entre imágenes generadas por modelos de tipo GAN y fotografías reales, definiendo explícitamente el problema como una tarea de clasificación binaria real vs. sintética [1]. Para ello, los autores construyen el conjunto de datos *Real or Synthetic Images* (RSI), que reúne imágenes sintéticas producidas por 12 arquitecturas GAN diferentes y un conjunto de imágenes reales emparejadas en cuanto a contenido y calidad visual, con el objetivo de favorecer la capacidad de generalización del detector frente a múltiples generadores [1, 23]. Sobre este conjunto de datos, evalúan varias CNN preentrenadas (entre ellas EfficientNetB4) utilizando aprendizaje por transferencia, y complementan el análisis con mapas de activación de clase (CAM) para identificar las regiones discriminativas que guían la decisión del modelo, reportando precisiones cercanas al 100 % en RSI y resultados competitivos en otros conjuntos de datos externos [1].

Metodológicamente, el trabajo es altamente comparable con el presente proyecto, ya que ambos formulan la detección de contenido generado por IA como una clasificación supervisada entre imágenes reales y sintéticas y se apoyan en CNN preentrenadas refinadas mediante aprendizaje por transferencia [1, 33, 34]. Sin embargo, existen diferencias importantes en el dominio de aplicación: mientras Baraheem y Nguyen utilizan un conjunto de datos genérico con imágenes sintetizadas por GAN, el presente trabajo se centra en un subconjunto de aproximadamente 10.000 imágenes noticiosas reales y sintéticas procedentes del conjunto MiRAGeNews, curado específicamente para el contexto de noticias en redes sociales [11]. Así, aunque las arquitecturas y el uso de transferencia son directamente reutilizables, la dificultad del problema cambia al considerar imágenes que incluyen texto incrustado, recortes propios de redes sociales y generadores modernos más allá de las GAN tradicionales, lo que hace que los resultados de RSI constituyan un techo optimista más que una estimación realista del rendimiento en el dominio noticioso [1].

En cuanto a limitaciones, el estudio se concentra en imágenes generadas por múltiples variantes de GAN y no discute explícitamente la robustez frente a modelos de difusión de última generación ni frente a generadores que no se hayan visto durante el entrenamiento [1]. Trabajos recientes sobre atribución de fuente y generalización abierta en detectores de imágenes sintéticas muestran que los modelos entrenados sobre un conjunto cerrado de generadores pueden degradar su desempeño de forma notable cuando se enfrentan a nuevos modelos o a cambios de concepto, como pasar de rostros a escenas, ilustraciones o imágenes panorámicas [3]. Esta brecha deja pendiente la evaluación sistemática del detector de Baraheem y Nguyen en escenarios de dominio abierto y en contextos específicos como imágenes noticiosas en redes sociales, donde la compresión, el recorte y la sobreimpresión de texto introducen artefactos adicionales. En respuesta a este vacío, el presente proyecto replantea la idea de entrenamiento/evaluación binaria hacia el dominio noticioso y compara varias arquitecturas modernas bajo un protocolo de benchmarking homogéneo en un contexto de redes sociales [3, 11].

2.4.2. Detection of AI-Generated Synthetic Images with a Lightweight CNN [2]

Lokner et al. proponen un enfoque de detección de imágenes sintéticas que prioriza la ligereza computacional mediante una CNN de solo ocho capas convolucionales y dos capas totalmente conectadas, diseñada para distinguir entre imágenes reales y sintéticas con un número reducido de parámetros y menores requisitos de memoria [2]. El estudio se centra en un contexto de observación de la Tierra: los autores construyen un conjunto de datos con imágenes satelitales Sentinel-2 reales y contrapartes sintéticas generadas con StyleGANv3, y además evalúan el modelo en dos conjuntos de referencia, CIFAKE y Midjourney v6, que contienen imágenes fotográficas y generadas por IA en múltiples categorías [2]. En términos de resultados, el modelo ligero propuesto alcanza una precisión de 97,32 % en CIFAKE, 99,94 % en Midjourney v6 y 98,69 % en el conjunto de entrenamiento satelital, superando o igualando a cuatro métodos de estado del arte más complejos, y complementa el análisis con técnicas de explicabilidad basadas en mapas de activación para mostrar que las activaciones en imágenes sintéticas tienden a concentrarse en bordes y texturas de fondo [2].

Aunque el dominio satelital difiere del escenario de imágenes noticiosas en redes sociales, el trabajo de Lokner et al. resulta altamente relevante para este proyecto en dos aspectos: por un lado, comparte el mismo planteamiento de clasificación binaria real vs. sintética con arquitecturas convolucionales y uso de conjuntos de datos con imágenes generadas por modelos modernos (StyleGANv3 y Midjourney); por otro, enfatiza explícitamente la relación entre complejidad del modelo y rendimiento, mostrando que es posible competir con arquitecturas profundas de gran tamaño mediante modelos más ligeros cuidadosamente diseñados [2]. Este énfasis es directamente utilizable en el presente trabajo de grado, donde se pretende comparar arquitecturas de aprendizaje profundo no solo en términos de métricas como exactitud, F1 o AUC-ROC, sino también en términos de eficiencia computacional y viabilidad de despliegue en entornos con recursos limitados, como plataformas de monitoreo de contenidos en redes sociales [39, 40]. De manera complementaria, el uso de métodos de explicabilidad en Lokner et al. refuerza la conveniencia de incluir análisis cualitativos

de mapas de activación o grad-CAM en este proyecto para entender qué patrones visuales utilizan los modelos al discriminar imágenes de noticias reales frente a sintéticas.

No obstante, la pertinencia del trabajo presenta varias limitaciones en relación con el problema planteado en este proyecto. En primer lugar, el conjunto de datos principal se basa en imágenes satelitales de Sentinel-2, con características espectrales y de resolución muy distintas a las de fotografías noticiosas compartidas en redes sociales; esto introduce un sesgo de dominio que dificulta extrapolar directamente los resultados de rendimiento al contexto noticioso [2]. En segundo lugar, aunque el modelo se evalúa en CIFAKE y Midjourney v6, el estudio no aborda explícitamente escenarios de dominio abierto en los que se pruebe con generadores no vistos durante el entrenamiento, aspecto que se ha identificado como crítico para la robustez de los detectores de imágenes sintéticas [3]. Finalmente, el trabajo se centra en la señal visual "limpia" y no considera artefactos característicos de la circulación en redes sociales, como compresión agresiva, reescalado, captura de pantalla o sobreimpresión de texto, que forman parte central del conjunto de datos MIRAGE-News utilizado en este proyecto [11]. En consecuencia, aunque Lokner et al. aportan evidencia de que modelos ligeros pueden competir en ciertos dominios, queda abierto cuán robustos resultan bajo distribuciones noticiosas con artefactos de red social; este vacío motiva evaluar arquitecturas bajo ese escenario y discutir eficiencia de manera comparable.

2.4.3. Methods and Trends in Detecting Generated Images: A Comprehensive Review [3]

La revisión de Mahara et al. sintetiza de forma sistemática los métodos contemporáneos para detectar y clasificar imágenes generadas por modelos de IA, con especial atención a contenido sintético producido por arquitecturas GAN y modelos de difusión [3]. Los autores proponen una taxonomía que agrupa los enfoques en detectores basados en CNN y transformers sobre el dominio espacial, métodos que explotan huellas en el dominio de la frecuencia o patrones de compresión, y técnicas de atribución de fuente que buscan identificar el generador responsable de una imagen sintética; además, se dedica una sección específica a marcos multimodales que combinan información visual con metadatos o texto asociado a la imagen [3]. La revisión también ofrece un inventario detallado de conjuntos de datos públicos para entrenamiento y evaluación, discutiendo sus sesgos de dominio (rostros, escenas, arte, satélite), la diversidad de generadores cubiertos y las implicaciones de estos factores sobre la capacidad de generalización de los detectores [3].

En cuanto a su comparabilidad con este proyecto, Mahara et al. formulan explícitamente el problema de la autenticidad de la imagen como una clasificación supervisada entre contenido real y generado, discuten métricas estándar como exactitud, F1 y AUC-ROC, y abogan por protocolos de benchmarking homogéneos, lo que coincide directamente con el planteamiento de comparar múltiples arquitecturas de aprendizaje profundo sobre un mismo conjunto de datos noticioso [3, 39]. Al mismo tiempo, la revisión pone de relieve limitaciones que este trabajo de grado puede abordar parcialmente, como la incapacidad de muchos detectores para generalizar a nuevos generadores o a nuevos tipos de contenido, la necesidad de evaluar escenarios de dominio abierto y la escasez de estudios centrados en imágenes de noticias o contextos de redes sociales frente al predominio

de rostros y arte digital en los conjuntos de datos existentes [3]. Tomando estas brechas como motivación, el presente proyecto se sitúa como un caso de estudio en el dominio noticioso y adopta un protocolo de comparación homogéneo para aportar evidencia empírica sobre rendimiento, trade-offs y limitaciones en un contexto de alta circulación y compresión en redes sociales.

2.4.4. ¿En qué se diferencia este proyecto de los trabajos mencionados?

La Tabla 2.1 sintetiza las diferencias metodológicas más relevantes entre los tres trabajos discutidos y el presente proyecto. La comparación se centra en el dominio de los datos, el tipo de generadores cubiertos, el objetivo experimental (benchmarking homogéneo vs. resultados dentro de un único dataset), y el nivel de discusión sobre generalización y despliegue.

En conjunto, el aporte diferencial del proyecto no radica en proponer un nuevo detector, sino en producir una comparación controlada y contemporánea de varios detectores en un dominio noticioso, explicitando la clase positiva (`fake`), el perfil de error y la sensibilidad a la elección de umbral, además de reconocer la variabilidad que surge al evaluar con conjuntos pequeños.

Dimensión	Baraheem & Nguyen [1]	Lokner et al. [2]	Mahara et al. [3]	Este proyecto
Tarea principal	Clasificación binaria real vs. sintética con CNN preentrenadas; énfasis en desempeño alto dentro del dataset RSI.	Clasificación binaria real vs. sintética con CNN ligera; énfasis en eficiencia y desempeño competitivo.	Revisión/taxonomía de métodos, datasets y tendencias; no propone un detector único.	Benchmarking comparativo de detectores de aprendizaje profundo (varias arquitecturas) bajo un protocolo común.
Dominio de datos	Dataset genérico (RSI), imágenes sintéticas de múltiples GANs y reales emparejadas.	Dominio satelital (Sentinel-2) + evaluación adicional en CIFAKE y Midjourney v6.	Multidominio (rostros, escenas, arte, satélite, etc.) según la literatura y datasets revisados.	Dominio noticioso/redes sociales: imágenes noticiables reales vs. clase `fake` (sintéticas) con artefactos típicos de circulación.
Generadores cubiertos	Conjunto cerrado de arquitecturas tipo GAN; cobertura limitada de difusión moderna.	StyleGANv3 y Midjourney (más evaluación que entrenamiento); cobertura dependiente de datasets usados.	Discute GAN y difusión, además de enfoques espaciales/frecuencia/compresión y atribución.	Entrenamiento/refinamiento en dominio noticioso y evaluación orientada a actualidad; se discute generalización frente a generadores no vistos como riesgo central.
Enfoque de evaluación	Métricas de clasificación y validación en RSI y algunos conjuntos externos; no se centra en incertidumbre o estabilidad con muestras pequeñas.	Métricas de clasificación; prioriza modelos pequeños y compara con métodos más complejos.	Recomienda buenas prácticas de benchmarking y métricas; identifica brechas (dominio abierto, sesgos).	Métricas interpretables con convención explícita de clase positiva `fake`; reporte de conteos (TP/FP/TN/FN) y énfasis en umbral/calibración e incertidumbre cuando n es pequeño.
Vacío que deja	Robustez frente a difusión y dominio abierto (generadores no vistos) y frente a contexto noticioso no caracterizado.	Transferibilidad a contexto noticioso y robustez a artefactos de redes sociales no evaluadas.	Identifica vacíos pero no entrega evidencia empírica específica para noticias en redes sociales.	Atiende parcialmente estos vacíos mediante evidencia empírica en un dominio noticioso y comparación homogénea entre detectores con discusión de trade-offs.

Cuadro 2.1: Comparación entre los trabajos relacionados discutidos y el presente proyecto.

Capítulo 3

Metodología

En este capítulo se presenta la metodología usada para el desarrollo de la comparativa de modelos de aprendizaje profundo con respecto a sus capacidades de distinguir entre imágenes noticiables reales e imágenes noticiables sintéticas, donde el contenido visual sintético se ha creado utilizando inteligencias artificiales generativas. El orden de esta metodología se constituye en las siguientes etapas: (1) selección de los modelos a utilizar en la comparativa, (2) selección del conjunto de datos para la aplicación de la técnica, (3) establecimiento de configuraciones para la aplicación de la técnica en cada modelo, y (4) diseño de la prueba de validación. La Figura 3.1 se plantea a continuación como apoyo visual para ilustrar el flujo de las etapas previamente descritas en mayor detalle.

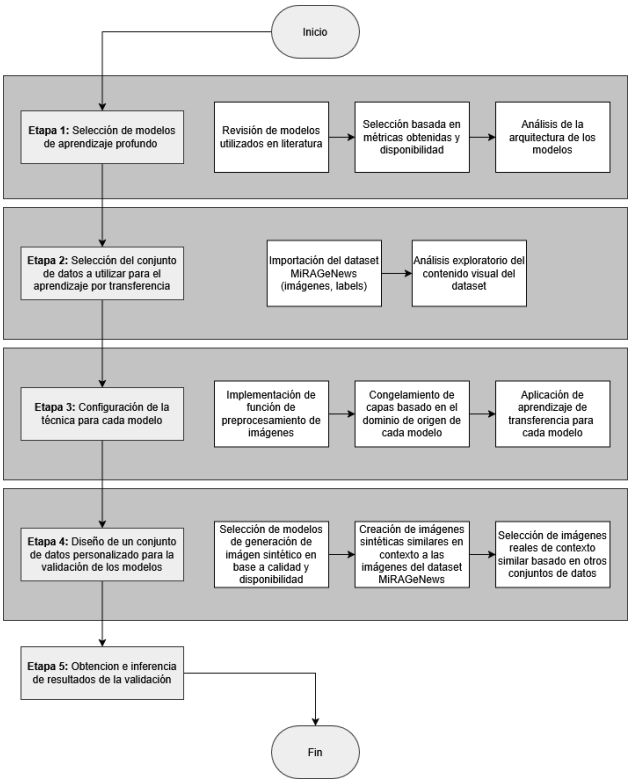


Figura 3.1: Esquema general del pipeline de la comparativa de los modelos de clasificación.

3.1. Selección de los modelos de aprendizaje profundo

En el marco del desarrollo y entrenamiento de modelos de aprendizaje profundo para clasificación de imágenes según su autenticidad visual en contexto noticioso, existe una amplia variedad de propuestas con enfoques y protocolos de evaluación distintos. En la práctica, no todos los modelos descritos en la literatura cuentan con código y pesos accesibles, y algunos repositorios asociados pueden estar incompletos o sujetos a restricciones de licencia. Por ello, esta comparativa selecciona modelos bajo criterios orientados a (i) mantener un respaldo técnico y metodológico suficiente, y (ii) garantizar que los modelos puedan ejecutarse y evaluarse bajo un protocolo homogéneo en el entorno experimental del proyecto. Los criterios base de selección se describen en la Tabla 3.2.

Criterio	Descripción
Rendimiento	El modelo debe haber demostrado un rendimiento competitivo en tareas de clasificación de imágenes reales vs sintéticas, con métricas reportadas que superen un umbral mínimo $\geq 80\,\%$ en conjuntos de datos relevantes.
Generalización	El modelo debe haber mostrado capacidad de generalización a generadores no vistos durante el entrenamiento, evidenciada por evaluaciones fuera de dominio o en benchmarks que incluyan múltiples familias de generadores.
Accesibilidad	El código fuente y los pesos preentrenados del modelo deben estar disponibles públicamente bajo una licencia que permita su uso para investigación y desarrollo, preferiblemente con documentación clara y ejemplos de uso.
Viabilidad práctica	El modelo debe ser factible de implementar y ejecutar en un entorno experimental con recursos computacionales limitados, considerando aspectos como la complejidad arquitectónica, los requisitos de hardware y el tiempo de inferencia.

Cuadro 3.2: Criterios de selección para los modelos de aprendizaje profundo a comparar.

Utilizando estos criterios como guía, se realizó una búsqueda de modelos de aprendizaje profundo aplicables a la detección de imágenes reales vs. sintéticas en escenarios abiertos. El conjunto final se organiza en dos grupos con roles distintos en la comparativa:

- **Modelos académicos de referencia:** propuestas descritas en literatura revisada por pares, con evaluación sistemática y énfasis explícito en generalización fuera de dominio.
- **Modelos abiertos de repositorios públicos:** detectores disponibles en Hugging Face seleccionados como *baselines* prácticos por su facilidad de integración y disponibilidad, sin asumir que sus métricas reportadas sean directamente comparables con benchmarks académicos.

Tras la revisión de literatura, se seleccionaron tres detectores académicos que, por sus resultados

reportados y su diseño, ofrecen un buen equilibrio entre rendimiento, generalización y viabilidad práctica:

- UniversalFakeDetect [7]
- DistilDIRE [8]
- D^3 (Discrepancy Deepfake Detector) [9]

UniversalFakeDetect se apoya en el espacio de características de CLIP (ViT-L/14) congelado y emplea clasificadores simples (vecinos más cercanos o una capa lineal) para distinguir imágenes reales de generadas, reportando mejoras sustanciales en exactitud al evaluar generadores de difusión y modelos autoregresivos no vistos, frente a detectores de referencia comparados por los autores [7]. De manera complementaria, DistilDIRE destila el comportamiento de un detector basado en reconstrucción con difusión (DIRE) en un modelo ligero que combina la imagen original con el ruido del primer paso de denoising, alcanzando en el benchmark DiffusionForensics precisiones cercanas o superiores al 98–99 % en generadores vistos como ADM y Stable Diffusion v1, y promedios alrededor del 89 % de exactitud y del 95 % de AP al considerar múltiples generadores en los subconjuntos tipo ImageNet y CelebA-HQ [8].

Por su parte, D^3 aborda el problema desde una perspectiva basada en discrepancias entre representaciones. Sea una imagen original

$$x,$$

y una versión transformada

$$\tilde{x} = T(x),$$

donde $T(\cdot)$ corresponde a una operación de corrupción basada en *patch shuffling*. El modelo extrae representaciones para ambas vistas,

$$z = f(x),$$

$$\tilde{z} = f(\tilde{x}),$$

y aprende patrones discriminativos a partir de la discrepancia entre dichas representaciones. Bajo esta formulación, el objetivo consiste en identificar artefactos persistentes asociados a imágenes sintéticas que permanecen observables incluso después de la perturbación estructural introducida por $T(\cdot)$. En un conjunto combinado de los datasets UniversalFakeDetect y GenImage que integra 20 generadores (8 vistos y 12 no vistos), el modelo reporta una exactitud media del 96.6 % en generadores dentro de su dominio y del 86.7 % en generadores fuera de dominio, con una exactitud global del 90.7 % y un AP del 96.8 %, superando en 5.3 puntos la exactitud media fuera de dominio de UFD y otros detectores de referencia como CNNDet, PatchFor, LNP y DIRE [9]. Estos resultados indican que los tres modelos seleccionados no solo muestran un rendimiento competitivo en

los benchmarks estándar, sino que además son robustos frente a variaciones arquitectónicas entre generadores y a cambios en el dominio visual.

En términos de accesibilidad, los tres modelos académicos ofrecen repositorios públicos con código y pesos que permiten reproducir su evaluación y reutilizar los componentes principales. UniversalFakeDetect publica los scripts necesarios para evaluar múltiples generadores y reentrenar la cabeza lineal sobre CLIP. DistilDIRE publica su implementación y pesos, y reporta una reducción importante de coste computacional (de 149.62 a 5.01 TFLOPs por imagen) y de tiempo de inferencia (factor aproximado de 3.2) respecto a DIRE, lo que lo hace viable en escenarios de alto volumen. Finalmente, D^3 también pone a disposición su código y pesos; aunque su arquitectura es más compleja que la de UFD o DistilDIRE, su diseño modular facilita adaptaciones por transferencia a dominios visuales específicos [7, 8, 9].

Finalmente, la potencial aplicabilidad de estos modelos en el contexto específico de detección de imágenes sintéticas asociadas a narrativas noticiosas usadas con fines de desinformación se basa en su diseño orientado a la generalización sobre generadores y dominios visuales diversos, y en su enfoque en artefactos de generación más que en el contenido semántico explícito. UniversalFakeDetect y D^3, al operar sobre características de modelos de visión-lenguaje preentrenados y ser entrenados con imágenes procedentes de un amplio espectro de generadores y conceptos, son candidatos naturales para adaptarse a dominios periodísticos mediante estrategias como el aprendizaje por transferencia o calibración sobre conjuntos de datos de noticias anotados. Por su lado, DistilDIRE ha demostrado un rendimiento elevado frente a motores de texto-a-imagen ampliamente utilizados en la práctica, como Stable Diffusion v2, DALL·E 2 e IF, lo que sugiere que puede desempeñar un papel relevante como detector especializado de contenido sintético en flujos de trabajo donde se sospeche la generación artificial de imágenes ilustrativas de noticias.

A nivel de licenciamiento de uso, los tres modelos académicos seleccionados (UniversalFakeDetect, DistilDIRE y D^3) se publican bajo licencias permisivas que permiten su uso para investigación y desarrollo, con código y pesos disponibles en repositorios públicos (GitHub). Más específicamente, UniversalFakeDetect se publica bajo una licencia MIT, DistilDIRE bajo una licencia Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0), y D^3 bajo una licencia Creative Commons Attribution 4.0 (CC BY 4.0). Estas licencias permiten la reutilización, modificación y distribución de los modelos, siempre que se atribuya adecuadamente a los autores originales y, en el caso de DistilDIRE, que no se utilice con fines comerciales. Esta disponibilidad y permisividad facilitan la integración de estos modelos en el marco experimental del proyecto, permitiendo su evaluación comparativa y posibles adaptaciones para el dominio noticioso.

Adicionalmente, y con el objetivo de aportar una dimensión aplicada, se incluyeron modelos disponibles en Hugging Face que pueden considerarse representativos de soluciones abiertas usadas en la práctica. En este proyecto, estos modelos no se tratan como equivalentes a los modelos académicos de referencia: su selección se apoya principalmente en disponibilidad, compatibilidad técnica y reproducibilidad (código/pesos públicos, licencia explícita para uso en investigación/desarrollo, documentación operativa y facilidad de integración con librerías estándar como Transformers), y sólo de forma secundaria en métricas reportadas por sus autores.

En particular, se seleccionaron tres detectores de Hugging Face por cubrir perfiles prácticos

distintos (generalista, humano-vs-IA de alta calidad y detector facial especializado):

- `Smogy/SMOGY-Ai-images-detector` [42]
- `Ateeqq/ai-vs-human-image-detector` [43]
- `prithivMLmods/deepfake-detector-model-v1` [44]

Los tres modelos seleccionados de Hugging Face comparten como objetivo original la detección binaria entre imágenes reales y sintéticas, aunque fueron concebidos para contextos distintos. `Smogy/SMOGY-Ai-images-detector` se diseñó como un detector generalista de imágenes generadas por IA, partiendo del modelo `Organika/sdxl-detector` y ajustándolo finamente con un conjunto heterogéneo de imágenes reales y sintéticas procedentes de fuentes abiertas. `Ateeqq/ai-vs-human-image-detector` se orienta a distinguir imágenes humanas frente a imágenes generadas por modelos de última generación (como Midjourney, Flux, Stable Diffusion 3.5 y GPT-4o), y se interpreta aquí como un baseline práctico para contenido sintético de alta calidad. Por su parte, `prithivMLmods/deepfake-detector-model-v1` es un detector especializado en deepfakes faciales; por lo tanto, su pertinencia para imágenes noticiosas generales es limitada y su comportamiento se analizará como referencia para escenarios donde la presencia de rostros sea dominante, no como un detector universal de noticias.

En cuanto a los datos de entrenamiento, `SMOGY-Ai-images-detector` se obtiene mediante el refinamiento del modelo `Organika/sdxl-detector` sobre un corpus heterogéneo de imágenes reales y generadas compiladas desde Reddit, conjuntos de Kaggle y arte de dominio público, integrando tanto contenido fotográfico como artístico. Esta diversidad de dominios busca capturar patrones generales de generación sintética más allá de un solo tipo de escena o estilo. Luego, el modelo `Ateeqq/ai-vs-human-image-detector` se entrena sobre un conjunto equilibrado de 60.000 imágenes generadas por IA y 60.000 imágenes humanas, incluyendo explícitamente muestras producidas por varios modelos generativos contemporáneos (Midjourney, Flux, Stable Diffusion 3.5, GPT-4o, entre otros), lo que introduce variabilidad en estilos, calidades y artefactos de generación. El modelo `deepfake-detector-model-v1` se basa en el dataset Deepfake-vs-Real-60K, un conjunto de aproximadamente 60.000 imágenes faciales de alta calidad que incluye en torno a 30.000 caras falsas (deepfake) y 30.000 caras reales, construido expresamente para entrenar y evaluar modelos de clasificación binaria en el contexto de detección de deepfakes; el conjunto mantiene un balance estricto de clases y está diseñado para tareas de entrenamiento, evaluación y benchmarking. Esta combinación de datos diversos, equilibrados y centrados en la dicotomía real/sintético proporciona una base sólida para transferir posteriormente estos modelos a escenarios donde se deban discriminar imágenes noticiosas auténticas frente a imágenes generadas.

En cuanto a rendimiento, las fichas técnicas de estos modelos reportan métricas elevadas en sus conjuntos de evaluación originales, superando el umbral de exactitud del 80 % considerado en el criterio de selección. No obstante, es importante matizar que estos valores no provienen necesariamente de benchmarks académicos homogéneos ni de protocolos revisados por pares, y por tanto se interpretan únicamente como indicadores preliminares de que los modelos funcionan en su

dominio de entrenamiento. En este proyecto, la evidencia comparativa se basará en re-evaluarlos bajo un mismo protocolo y un mismo dominio noticioso.

Con esa advertencia, los resultados reportados por sus autores incluyen: (i) para `Smogy/SMOGY-Ai-images-detector`, exactitud 0,9818 y métricas cercanas a 98 % en el split de prueba del conjunto base, además de resultados fuera de dominio entre 75 % y 91 % según el generador [42]; (ii) para `Ateeqq/ai-vs-human-image-detector`, exactitud de 99.23 % y F1 ponderado de 99.23 % [43]; y (iii) para `prithivMLmods/deepfake-detector-model-v1`, exactitud global de 94.44 % y F1 superiores al 94 % en un conjunto facial de prueba [44].

En términos de accesibilidad, los tres modelos destacan por la disponibilidad de código y pesos preentrenados en Hugging Face. `SMOGY-Ai-images-detector` hereda la estructura del modelo en el que se basó y ofrece pesos listos para usar además de variantes cuantizadas, lo que reduce la barrera de entrada para su despliegue en entornos experimentales y de producción. Similarmente, `Ateeqq/ai-vs-human-image-detector` incluye instrucciones explícitas de instalación vía pip y ejemplos de uso, lo que evidencia una comunidad activa y facilita su integración en flujos de trabajo diversos. Además de proporcionar código de inferencia con un tutorial web explicando cómo se refinó el modelo, `deepfake-detector-model-v1` también provee una interfaz para experimentación y un conjunto de datos asociado (Deepfake-vs-Real-60K) disponible en Hugging Face, lo que permite reproducir y extender los experimentos originales con relativa facilidad. En contraste, parte de la literatura especializada en detección de deepfakes describe modelos de alto rendimiento cuyo código o pesos no se encuentran plenamente disponibles, o que requieren configuraciones complejas de hardware y software, lo que refuerza el valor de estos tres modelos por su enfoque abiertamente reproducible y su baja fricción de adopción.

Teniendo en cuenta el licenciamiento de uso, los tres modelos de Hugging Face seleccionados se publican bajo licencias permisivas que permiten su uso para investigación y desarrollo, con código y pesos disponibles públicamente. El modelo `SMOGY-Ai-images-detector` se publica bajo una licencia Creative Commons Attribution Non Commercial 4.0 (CC BY-NC 4.0) [42], lo que permite su uso, modificación y distribución para fines no comerciales, siempre que se atribuya adecuadamente a los autores originales. Sumado a esto, tanto `ai-vs-human-image-detector` como `deepfake-detector-model-v1` se publican bajo la licencia Apache 2.0 [43, 44], que es una licencia de software permisiva que permite el uso, modificación y distribución del software, incluso para fines comerciales, siempre que se cumplan ciertas condiciones como la inclusión de un aviso de derechos de autor y la renuncia de garantías. Estas licencias facilitan la integración de estos modelos en el marco experimental del proyecto, permitiendo su evaluación comparativa y posibles adaptaciones para el dominio noticioso sin restricciones significativas.

Considerando el contexto específico de la verificación de imágenes noticiosas, estos modelos abiertos se incorporan con una expectativa de pertinencia heterogénea. `SMOGY-Ai-images-detector` es el candidato más alineado con el objetivo general (real vs. sintético en escenas diversas) por su planteamiento generalista. `Ateeqq/ai-vs-human-image-detector` se considera un baseline práctico útil para estimar el desempeño de un detector orientado a contenido sintético de alta calidad, aunque su dominio de entrenamiento puede no coincidir con la distribución visual de imágenes periodísticas. Finalmente, `prithivMLmods/deepfake-detector-model-v1` se incorpora como baseli-

ne especializado para contenido con rostros, con la advertencia de que puede degradar su cobertura cuando la imagen no está dominada por señales faciales.

Integrados en un mismo marco experimental junto con los modelos seleccionados desde la literatura académica, estos detectores de Hugging Face permiten contrastar dos realidades complementarias: (i) el rendimiento de modelos académicos diseñados explícitamente para generalizar a generadores no vistos, y (ii) el desempeño de soluciones abiertas listas para usar, que priorizan accesibilidad y despliegue rápido pero pueden estar especializadas o calibradas a dominios distintos. Esta distinción es clave para una comparación honesta y para interpretar correctamente las limitaciones y alcances de cada enfoque en el dominio noticioso.

3.1.1. Arquitectura del modelo UniversalFakeDetect

UniversalFakeDetect (UFD) es un detector de imágenes sintéticas que reutiliza la red troncal de CLIP (ViT-L/14) como extractor de características y aprende un clasificador simple sobre ese espacio de representación. La idea central es que, en lugar de entrenar desde cero una CNN binaria real/falso, se puede explotar un espacio multimodal ya preentrenado para obtener un detector competitivo y con buen comportamiento fuera de dominio.

El trabajo parte de la observación de que una "receta estándar" (entrenar una red profunda de extremo a extremo para clasificar imágenes como reales o generadas por un conjunto limitado de generadores) tiende a degradarse al evaluar imágenes de familias no vistas, como modelos de difusión. Una explicación plausible es que el clasificador aprende artefactos específicos de los generadores vistos; en consecuencia, la clase "real" puede comportarse como un sumidero que absorbe imágenes sintéticas que no comparten exactamente esas huellas.

UFD propone un enfoque distinto: en lugar de entrenar una red completa para resolver directamente la clasificación binaria, utiliza el espacio de representación aprendido por CLIP como extractor fijo de características.

Sea una imagen de entrada $x \in \mathbb{R}^{H\times W\times C}$ y sea

$$f_{\text{CLIP}} : \mathbb{R}^{H\times W\times C} \to \mathbb{R}^d$$

la función de extracción de características implementada por el backbone ViT-L/14. El embedding asociado a la imagen se define como

$$z = f_{\text{CLIP}}(x),$$

donde $z \in \mathbb{R}^d$ corresponde a la representación latente utilizada por el detector.

La tarea de clasificación consiste entonces en aprender una función

$$g : \mathbb{R}^d \to \{0, 1\},$$

donde la etiqueta 0 representa imágenes reales y la etiqueta 1 representa imágenes sintéticas. Bajo esta formulación, el problema de detección se transforma en una tarea de clasificación sobre el

espacio de embeddings generado por CLIP, manteniendo congelados los parámetros del backbone durante todo el proceso de entrenamiento.

La Figura 3.2 ilustra el esquema de procesamiento cuando se utiliza un clasificador por vecinos más próximos.

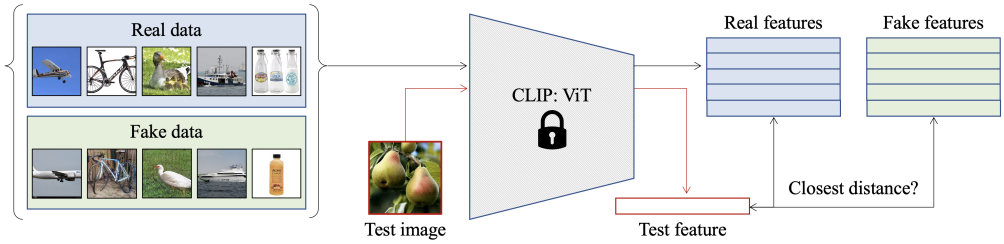


Figura 3.2: Esquema de procesamiento de la arquitectura de UniversalFakeDetect con un clasificador por vecinos más próximos (Fuente: [7]).

En el flujo de inferencia, una imagen de entrada se normaliza y reescala según los requisitos del backbone CLIP, y se alimenta al transformador ViT-L/14 para obtener un vector de características de alta dimensión. Este vector constituye la representación fija sobre la que opera UniversalFakeDetect; no se actualizan los parámetros de CLIP durante el entrenamiento del detector, de modo que el backbone actúa como extractor genérico de semántica y estilo visual. Sobre estas características, se exploran dos instancias de detector:

- Un **clasificador lineal**, que recibe como entrada el embedding z producido por CLIP y calcula una puntuación

$$s = Wz + b,$$

donde W y b corresponden a los parámetros entrenables de la cabeza de clasificación. Posteriormente, la probabilidad de pertenecer a la clase sintética se estima mediante

$$p(y = 1|x) = \sigma(s),$$

siendo $\sigma(\cdot)$ la función sigmoide. Durante el refinamiento únicamente se actualizan los parámetros W y b, mientras que el backbone CLIP permanece congelado.

- Un **clasificador k-NN**, que implementa una estrategia no paramétrica. Dado un embedding de consulta z_q, se calcula su distancia respecto a los embeddings de entrenamiento

$$\mathcal{D} = \{z_1, z_2, \ldots, z_n\},$$

utilizando una métrica de similitud o distancia (por ejemplo, distancia euclidiana o similitud coseno). Posteriormente se selecciona el conjunto de los k vecinos más próximos

$$N_k(z_q),$$

y la etiqueta predicha se determina mediante votación mayoritaria sobre las etiquetas asociadas a dichos vecinos.

En ambos casos, el diseño busca reducir el riesgo de aprender artefactos demasiado específicos de un generador concreto, ya que el espacio de representación subyacente fue optimizado para correspondencia imagen-texto y no para separar directamente real/falso. En consecuencia, el método apunta a mejorar el rendimiento fuera de dominio (OOD) frente a detectores tradicionales entrenados de extremo a extremo.

Aplicación en este proyecto.

- **Entrada:** imagen en su tamaño/resolución original; el preprocesamiento del repositorio normaliza y reescala según los requisitos de CLIP.
- **Congelado:** backbone CLIP (ViT-L/14) como extractor de características.
- **Refinado:** capa de salida de clasificación (cabeza k-NN real/fake).
- **Salida:** un puntaje (logit/probabilidad) para la clase, usado para construir la decisión binaria y las métricas reportadas.

3.1.2. Arquitectura del modelo DistilDIRE

En el contexto de distinguir la autenticidad de imágenes noticiosas, el método DIRE (*Diffusion Reconstruction Error*) muestra que se puede explotar la información implícita de un modelo de difusión sobre la distribución de imágenes para detectar contenido sintético. Sin embargo, DIRE es costoso porque requiere pasos de difusión y reconstrucción por imagen, lo que limita su uso a gran escala.

DistilDIRE se propone como una alternativa más ligera: intenta conservar parte de la capacidad de detección de DIRE, reduciendo el coste computacional mediante un esquema de destilación de conocimiento.

DistilDIRE es un detector binario que combina señales derivadas de un modelo de difusión con un modelo discriminativo (ResNet-50) para decidir si una imagen es real o sintética (por difusión o por GANs). En términos generales, busca aproximar el comportamiento de DIRE sin ejecutar una reconstrucción completa en inferencia, con mejoras reportadas en coste y tiempo por imagen.

La idea central es destilar en un modelo más ligero (estudiante) información asociada tanto a un modelo de difusión preentrenado como a una ResNet-50 preentrenada, de modo que el detector capture señales útiles sin incurrir en el coste completo del pipeline de reconstrucción.

Como profesor, se emplea una ResNet-50 preentrenada (ImageNet-1K) congelada durante el entrenamiento, de la cual se toman mapas de características antes de la cabeza clasificadora como representaciones de alto nivel. El modelo estudiante se entrena sobre una entrada extendida y

produce una salida binaria real/falso. Sobre sus representaciones internas se impone una pérdida de distilación (p.ej., MSE) para aproximar las características del profesor, además de una pérdida de clasificación (p.ej., BCE) respecto a las etiquetas [8].

La pérdida total combina ambos términos con un hiperparámetro λ ajustado por validación. El esquema general de este proceso se ilustra en la Figura 3.3.

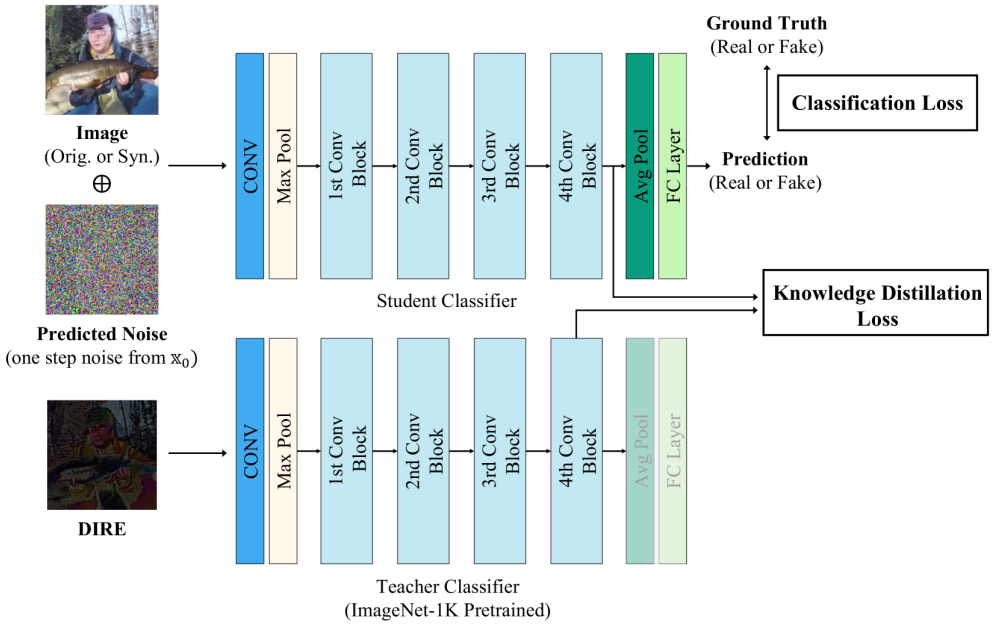


Figura 3.3: Esquema del framework de la arquitectura de DistilDIRE (Fuente: [8]).

El pipeline de DistilDIRE puede describirse formalmente como una composición de tres etapas: extracción de ruido latente, construcción de la representación de entrada y clasificación binaria.

Sea $x_0 \in \mathbb{R}^{H\times W\times 3}$ una imagen RGB de entrada. Mediante el procedimiento de inversión DDIM se estima una aproximación al ruido inicial del proceso generativo,

$$\varepsilon_0 = \mathrm{DDIMInv}(x_0),$$

donde ε_0 representa la señal latente asociada al modelo de difusión preentrenado.

Posteriormente, ambas representaciones se concatenan a nivel de canales,

$$z = [x_0; \varepsilon_0],$$

obteniendo una representación enriquecida que combina información visual observable e información relacionada con la dinámica inversa del proceso de difusión.

La representación resultante es procesada por el modelo estudiante basado en ResNet-50,

$$h_s = f_{\theta_s}(z),$$

mientras que la imagen original es procesada por la red maestra congelada,

$$h_t = f_{\theta_t}(x_0),$$

donde θ_t permanece fijo durante el entrenamiento.

Finalmente, la salida del clasificador se obtiene mediante una capa lineal seguida de una función sigmoidal,

$$\hat{y} = \sigma(W h_s + b),$$

produciendo una probabilidad de pertenencia a la clase sintética.

Aunque el entrenamiento original se centra en conjuntos de datos genéricos, el diseño de DistilDIRE puede ser relevante para imágenes noticiosas en la medida en que combina una señal generativa (ruido asociado a difusión) con representaciones discriminativas (ResNet). No obstante, su desempeño puede degradarse si el prior del difusor o el protocolo de extracción de ruido no están alineados con el dominio o con los generadores presentes.

Aplicación en este proyecto. La implementación utilizada conserva la estructura propuesta por los autores originales. Durante la fase de preprocesamiento se ejecuta la inversión DDIM para cada imagen del conjunto de entrenamiento con el fin de obtener la representación ε_0. Este procedimiento se realiza una única vez y sus resultados son almacenados para evitar recomputación durante las épocas posteriores.

- **Entrada:** par (x_0, ε_0) generado a partir de cada imagen del conjunto MiRAGeNews.
- **Congelado:** modelo de difusión utilizado para la inversión DDIM y backbone ResNet-50 preentrenado.
- **Refinado:** parámetros de la cabeza de clasificación y componentes específicos del estudiante definidos por el repositorio original.
- **Optimización:** actualización mediante Adam utilizando la función híbrida de clasificación y destilación.
- **Salida:** probabilidad $\hat{y} \in [0, 1]$ asociada a la clase sintética.

Bajo esta configuración, el aprendizaje por transferencia se restringe exclusivamente a los parámetros de decisión final, preservando las representaciones aprendidas durante el preentrenamiento y reduciendo el riesgo de sobreajuste.

3.1.3. Arquitectura del modelo D^3

El modelo D^3 (*Discrepancy Deepfake Detector*) es un detector basado en CLIP que introduce explícitamente una señal de discrepancia entre una imagen y una versión fuertemente distorsionada para aprender artefactos compartidos por múltiples generadores. El objetivo es mejorar la generalización frente a generadores no vistos y degradaciones del mundo real.

Su contribución central es ir hacia un escenario multi-generador: entrenar y evaluar sobre múltiples familias (GAN y difusión), de modo que el detector aprenda patrones más estables y menos dependientes de un único generador. El "truco" metodológico consiste en explotar la discrepancia entre dos vistas correlacionadas (original vs distorsionada), lo que busca reducir el riesgo de sobreajuste a huellas particulares y enfatizar señales invariantes.

El modelo recibe imágenes RGB con etiqueta binaria (real/generada). Durante el entrenamiento, cada ejemplo produce dos vistas: (i) la imagen "original" tras aumentos estándar (p.ej., redimensionamiento, recortes, desenfoque, compresión), y (ii) una imagen distorsionada obtenida aplicando *patch shuffling*, que destruye la estructura semántica global pero preserva señales de bajo nivel.

Dadas las dos vistas de entrada,

$$x \quad \text{y} \quad \tilde{x} = T(x),$$

ambas son procesadas por el mismo backbone CLIP congelado,

$$h_1 = f_{\text{CLIP}}(x),$$

$$h_2 = f_{\text{CLIP}}(\tilde{x}),$$

obteniendo dos representaciones en el espacio de características compartido.

Posteriormente, estas representaciones son fusionadas mediante un bloque de autoatención parametrizado por θ_{att},

$$h_f = A_{\theta_{att}}(h_1, h_2),$$

donde el mecanismo de atención aprende relaciones de consistencia y discrepancia entre ambas vistas.

La representación fusionada se entrega finalmente a una capa completamente conectada,

$$\hat{y} = W h_f + b,$$

que genera la decisión binaria real/sintética.

Bajo esta formulación, el conjunto de parámetros optimizados durante el refinamiento queda definido como

$$\theta = \{\theta_{att}, W, b\},$$

manteniendo congelados todos los parámetros asociados al backbone CLIP.

La Figura 3.4 muestra el esquema de la arquitectura.

En el ámbito noticioso, las imágenes sintéticas se generan con una variedad de modelos (a menudo desconocidos al momento del despliegue). Un detector entrenado sólo con un generador corre el riesgo de degradarse cuando aparecen nuevas arquitecturas o pipelines. D^3 intenta mitigar esta fragilidad al aprender señales a partir de la discrepancia entre vistas, lo que puede favorecer patrones más estables frente a cambios de dominio.

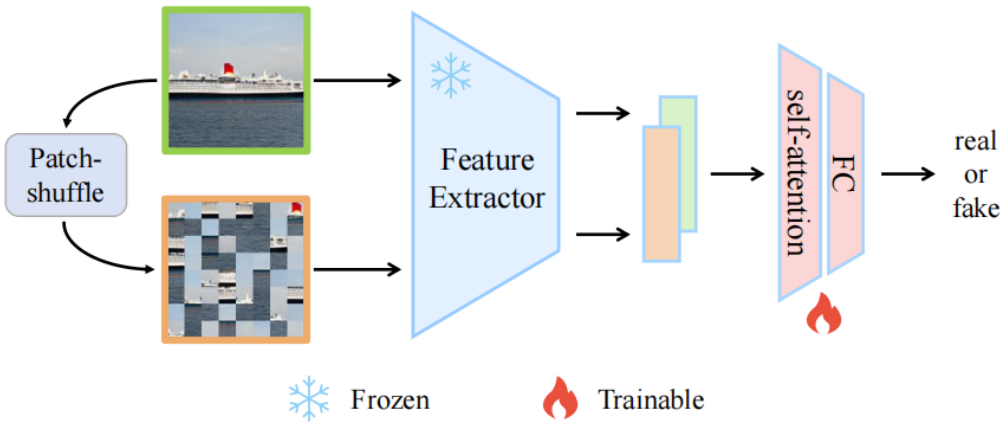


Figura 3.4: Esquema del framework de la arquitectura de D^3 (Fuente: [9]).

Aplicación en este proyecto.

- **Entrada:** imagen en su tamaño/resolución original como punto de partida; el preprocesamiento del repositorio genera las dos vistas (aumentada y distorsionada) requeridas por el modelo.
- **Congelado:** backbone CLIP.
- **Refinado:** módulos cercanos a la salida (autoatención de fusión y capa totalmente conectada final).
- **Salida:** un puntaje (logit/probabilidad) para la clase.

3.1.4. Arquitecturas de los modelos extraídos de repositorios públicos

Los modelos extraídos de repositorios de Hugging Face difieren en la arquitectura base que emplean, aunque comparten un patrón general: un backbone tipo Vision Transformer y una cabeza de clasificación binaria.

En el caso de `Smogy/SMOGY-Ai-images-detector`, se utiliza Swin Transformer (*Shifted Window Transformer*), una arquitectura tipo ViT diseñada como backbone generalista para múltiples tareas. A diferencia de un Transformer visual "plano" (una única cuadrícula de patches con resolución fija), Swin introduce una estructura por etapas con resolución decreciente y canales crecientes, análoga al comportamiento jerárquico de redes como ResNet.

El componente central de Swin es el mecanismo de "ventanas desplazadas" (shifted windows): los bloques se alternan entre una configuración de autoatención sobre ventanas fijas (W-MSA) y una configuración en la que dichas ventanas se desplazan cíclicamente, lo que introduce conexiones entre regiones que en el bloque anterior estaban separadas. Esta alternancia permite que la red modele interacciones de largo alcance sin recurrir a autoatención global, preservando la eficiencia del cálculo local al tiempo que se genera una representación jerárquica que captura tanto detalles

locales como estructuras globales. La Figura 3.5 muestra el esquema general de la arquitectura del Swin Transformer.

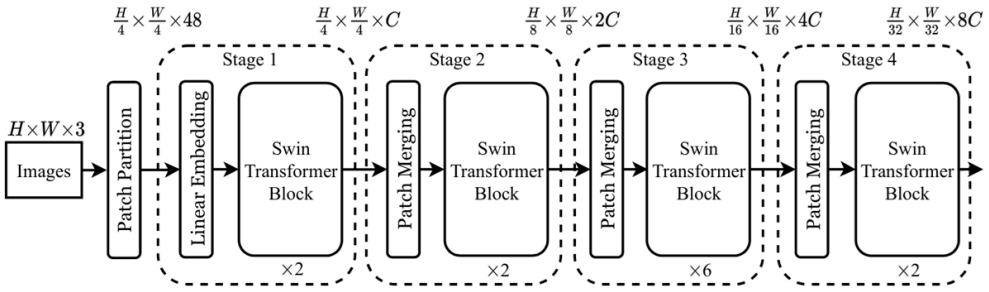


Figura 3.5: Esquema general de la arquitectura del Swin Transformer (Fuente: [10]).

Swin recibe como entrada una imagen RGB de tamaño fijo que se tokeniza mediante una capa de "patch embedding", convirtiendo bloques espaciales de la imagen en vectores de dimensión d. Posteriormente, sucesivas etapas de bloques Swin Transformer aplican autoatención local en ventanas, seguidas de bloques de MLP y operaciones de "patch merging" que reducen la resolución espacial mientras incrementan la dimensión de los canales, produciendo un mapa de características jerárquico. Al final del backbone, las características agregadas se proyectan mediante pooling global a un espacio de salida de baja dimensión que alimenta una cabeza de clasificación lineal para predecir las etiquetas de la tarea objetivo.

Por otro lado, `Ateeqq/ai-vs-human-image-detector` y `prithivMLmods/deepfake-detector-model-v1` se basan en la arquitectura SigLIP como backbone visual. SigLIP se presenta como un marco de preentrenamiento multimodal en el que un codificador de imágenes y un codificador de texto se entrenan conjuntamente mediante una pérdida sigmoide pareja para alinear representaciones visuales y lingüísticas. La clave metodológica de SigLIP reside en sustituir la pérdida contrastiva clásica basada en softmax global (como en CLIP) por una pérdida sigmoide aplicada de forma independiente a cada par imagen-texto, lo que elimina la necesidad de normalizar sobre todas las combinaciones de pares del batch.

El codificador visual de SigLIP recibe como entrada una imagen RGB (a partir de la cual se aplica el redimensionamiento requerido por la arquitectura), la tokeniza en patches y proyecta cada uno a un vector de dimensión d para formar una secuencia de entrada del Transformer. A lo largo de las capas, la autoatención multicabeza permite que cada patch agregue información de otros patches, capturando relaciones globales en la escena, mientras que los bloques MLP y las normalizaciones de capa refinan la representación.

3.2. Selección del conjunto de datos para el entrenamiento de los modelos

En el marco de conjuntos de datos orientados a la clasificación de imágenes sintéticas, se ha identificado el dataset MiRAGeNews [11] como una opción particularmente apta para adaptar (por aprendizaje por transferencia) los modelos seleccionados al dominio noticioso. MiRAGeNews es un dataset multimodal diseñado para el estudio y la detección de noticias falsas generadas mediante modelos de difusión, combinando imágenes hiperrealistas y pies de foto potencialmente dañinos o engañosos. El conjunto de datos se propone como banco de pruebas para métodos de detección de contenido noticioso generado por IA, en un escenario donde tanto la imagen como el texto pueden ser sintéticos.

En comparación con datasets previos de desinformación noticiosa, MiRAGeNews se centra en formas particularmente peligrosas de desinformación: noticias con imágenes fotorealistas completamente generadas por modelos de difusión y textos diseñados para ser inflamatorios o misinformativos. Por esto, el conjunto de datos se alinea de manera directa con el objetivo del proyecto de evaluar la capacidad de los modelos para discriminar entre imágenes noticiosas reales y sintéticas, especialmente en un contexto donde la generación de contenido visual se ha vuelto cada vez más sofisticada y accesible.

Aunque el dataset incluye *train*, *validation* y varios splits de *test*, en este trabajo el refinamiento se realiza exclusivamente con las 10.000 imágenes del split *train* (NYT como real y Midjourney v5.2 como sintético), manteniendo el balance 50 %/50 % entre clases. Los splits *validation* y *test* de MiRAGeNews se utilizan únicamente como referencia y para análisis exploratorio de calidad (duplicados, solapamientos y propiedades de imagen), pero no como evaluación principal ni como fuente de selección de hiperparámetros. En términos de pertinencia como conjunto de datos para el refinamiento visual, MiRAGeNews se escoge por la procedencia real desde fotoperiodismo (NYT/BBC/CNN) que aporta escenas y estética representativas del dominio noticioso, la generación sintética orientada explícitamente a imitar fotografía de prensa (*news photo style*) con modelos de difusión hiperrealistas, y finalmente el balance de clases que permite estimar desempeño sin depender de desbalances artificiales.

3.2.1. Procedencia del conjunto de datos

La parte real del dataset se deriva de TARA, un corpus de noticias del New York Times (NYT) que contiene pares imagen-subtítulo anotados con información detallada de tiempo y lugar del evento noticioso. Los autores muestrean 6.500 pares de imagen y pie de foto de TARA y los consideran como el subconjunto de "noticias reales", asegurando que cada caption incluya referencias geográficas y temporales que constriñen la generación posterior. Para la evaluación fuera de dominio, se recolectan además 250 pares imagen-subtítulo de BBC y 250 de CNN, procedentes de un dataset de artículos en Kaggle, que actúan como fuentes reales adicionales con estilos editoriales diferentes a NYT.

Luego, a partir de cada subtítulo real, se utiliza GPT-4 para producir una versión ficticia que

pueda considerarse "dañina o engañosa", manteniendo las entidades nombradas del texto original para no alejarse demasiado del contenido factual de partida. El prompt explícito pide pies de foto ficticios que podrían considerarse "dañinos o engañosos", lo que sitúa el dataset en un contexto de desinformación deliberada más que en simples errores benignos.

En el conjunto dentro de dominio (entrenamiento y validación), las imágenes falsas se generan con Midjourney v5.2, seleccionado por su capacidad de producir imágenes hiperrealistas y por una moderación relativamente laxa respecto a contenidos sensibles. Para guiar la generación se utilizan como prompt los captions falsos de GPT-4, junto con palabras clave que fijan el estilo "news photo style", el aspecto de fotografía de prensa y la relación de aspecto tomada de la imagen real correspondiente. Para las particiones de prueba fuera de dominio, las imágenes falsas no se crean con Midjourney sino con DALL-E 3 y Stable Diffusion XL (SDXL), combinados con captions de BBC y CNN, generando así cuatro subconjuntos OOD: BBC+DALL-E 3, CNN+DALL-E 3, BBC+SDXL y CNN+SDXL.

3.2.2. Descripción general del conjunto de datos

Las imágenes reales provienen de fotoperiodismo profesional de NYT, BBC y CNN, con composición típica de prensa: escenas de protestas, mítines políticos, catástrofes, retratos de líderes y tomas institucionales. Estas imágenes incorporan variación natural en condiciones de iluminación, ángulos de cámara, densidad de personas y contexto urbano o rural, reflejando la estética habitual de la fotografía de noticias.

Las imágenes falsas dentro de dominio son generadas exclusivamente con Midjourney v5.2, que produce imágenes hiperrealistas con alta fidelidad de texturas, iluminación y composición, hasta el punto de resultar difíciles de distinguir. En las particiones fuera de dominio, las imágenes sintéticas se generan con DALL-E 3 y Stable Diffusion XL, que tienden a producir resultados algo menos convincentes que Midjourney según los ejemplos proporcionados, pero siguen siendo suficientemente realistas para desafiar a modelos automáticos y observadores humanos. El conjunto completo cubre, por tanto, varias familias de modelos de difusión, lo que permite estudiar la generalización de detectores entre generadores. En la Figura 3.6, se presentan dos ejemplos para cada split de imágenes que se pueden hallar en el conjunto de datos, donde el par superior corresponden a imágenes falsas, y el par inferior corresponden a imágenes reales.

En términos de distribución, el conjunto de datos está formado por 12.500 pares para entrenamiento y validación, junto con 2.500 pares para prueba, totalizando 15.000 ejemplos multimodales. En la literatura asociada se detalla la división: 10.000 pares para entrenamiento, 2.500 para validación y cinco subconjuntos de test de 500 pares cada uno, donde cuatro son fuera de dominio (BBC/CNN con DALL-E 3 o SDXL) y uno es dentro de dominio (NYT & Midjourney) [11]. Además, en todas las particiones, la tarea está formulada como clasificación binaria de pares imagen-caption: etiqueta 0 cuando tanto la imagen como el texto son reales, y etiqueta 1 cuando ambos son generados.

Además, el balance de dicha distribución es equilibrado. En el conjunto de entrenamiento y validación, los autores parten de 6.250 pares reales de NYT y generan 6.250 pares falsos adicionales

Figura 3.6: Ejemplos de imágenes del conjunto de datos MiRAGeNews (Fuente: [11]).

con GPT-4 y Midjourney, alcanzando los 12.500 pares mencionados. La partición de prueba de 2.500 pares se construye con 500 pares dentro de dominio (NYT & Midjourney) y 2.000 pares fuera de dominio, de manera que cada uno de los cinco subconjuntos de 500 ejemplos está balanceado entre pares reales y falsos, sumando a 1.250 instancias de cada clase, con 250 de cada combinación dentro y fuera de dominio.

3.2.3. Limitaciones del conjunto de datos

En el ámbito técnico del contenido de MiRAGeNews, este está fuertemente influenciado por las características de Midjourney v5.2, que constituye la fuente principal de imágenes generadas en el conjunto dentro de dominio. Aunque esto permite evaluar la capacidad de los modelos para detectar imágenes sintéticas de alta calidad, también introduce un sesgo hacia las características específicas de este modelo, lo que puede limitar la generalización a otras familias de generadores. Además, los modelos de difusión empleados tienen sus propios sesgos en cuanto a distribución demográfica, estética y representación de grupos sociales, heredados de sus datos de entrenamiento, que se reflejan indirectamente en el dataset. Estos sesgos pueden manifestarse en la forma en que se representan género, raza, vestimenta, violencia o roles de poder en las imágenes generadas.

MiRAGeNews define la etiqueta a nivel de par imagen-subtítulo: etiqueta 0 cuando tanto la

imagen como el texto son reales, y etiqueta 1 cuando ambos son generados. Dado que el objetivo de este proyecto es evaluar detectores visuales, se usa únicamente la imagen y se toma la etiqueta del par como proxy de autenticidad visual. Este supuesto es razonable en la medida en que la construcción del dataset separa fuentes de imagen (fotoperiodismo vs. generadores) y genera la imagen sintética condicionada por el caption ficticio; no obstante, metodológicamente implica que el modelo podría estar aprendiendo a discriminar un pipeline de generación (fuente/generador y su estética) más que una noción general de autenticidad visual. Por ello, los resultados se interpretan como evidencia de transferencia al dominio noticioso bajo el protocolo de MiRAGeNews y se complementan con una evaluación externa (conjunto de validación personalizado).

En los metadatos públicos del recurso no se identifica una licencia explícita y unificada para MiRAGeNews. En consecuencia, su utilización en este proyecto se delimita a fines académicos, bajo las siguientes condiciones operativas: (i) citación explícita del dataset y su literatura asociada, (ii) no redistribución del dataset ni publicación de derivados que lo reempaqueten, y (iii) uso limitado a entrenamiento/evaluación internos, respetando los términos aplicables de las fuentes subyacentes y de las herramientas de generación involucradas [11].

3.2.4. Análisis exploratorio de los datos del conjunto de datos

Se realizó un análisis exploratorio del conjunto MiRAGeNews con el objetivo de (i) verificar la distribución y consistencia de las particiones, (ii) detectar duplicados y posibles fugas entre splits (que inflan métricas y comprometen validez), y (iii) caracterizar propiedades de las imágenes (resolución y formato) que puedan inducir atajos ajenos al fenómeno de interés (autenticidad visual). Dado que el proyecto trabaja únicamente con la modalidad de imagen, este EDA se enfoca en señales visuales y en la integridad del protocolo de particionado.

La Tabla 3.3 resume la procedencia y el tamaño de cada partición (reportado como número de pares imagen-subtítulo, equivalente al número de imágenes). Se reporta explícitamente el balance de clases (real vs fake) por split, ya que el desbalance altera la interpretabilidad de métricas agregadas (accuracy) y puede inducir umbrales subóptimos en evaluación.

Split	Fuente Imagen	Tamaño	Balance de Clases
Train	NYT (real) / Midjourney (fake)	10,000 pares	50 % real / 50 % fake
Validation	NYT (real) / Midjourney (fake)	2,500 pares	50 % real / 50 % fake
Test 1	NYT (real) / Midjourney (fake)	500 pares	50 % real / 50 % fake
Test 2	BBC (real) / DALL-E 3 (fake)	500 pares	50 % real / 50 % fake
Test 3	CNN (real) / DALL-E 3 (fake)	500 pares	50 % real / 50 % fake
Test 4	BBC (real) / SDXL (fake)	500 pares	50 % real / 50 % fake
Test 5	CNN (real) / SDXL (fake)	500 pares	50 % real / 50 % fake

Cuadro 3.3: Resumen de las particiones del conjunto de datos MiRAGeNews.

Para comprobar la limpieza concreta de este dataset, resulta necesario verificar tanto la existencia de duplicados, como posibles fugas de información. La presencia de ambos entre entrenamiento y

prueba sesga las métricas al alza y compromete la validez experimental. Para evaluar la posibilidad de fuga de información entre particiones, se buscaron duplicados exactos y near-duplicados: (i) duplicados exactos mediante hashing determinista del contenido de la imagen, y (ii) cuasi-duplicados mediante hashing perceptual (pHash/aHash) y umbrales de distancia de Hamming, con un umbral de 5. Se reporta el número de colisiones entre splits (que puede inflar el "tamaño efectivo" del dataset) y el solapamiento entre train/validation y cada split de prueba en las figuras 3.7 y 3.8.

En particular, se definió un cuasi-duplicado como un par de imágenes con distancia de Hamming ≤ 5 sobre un hash perceptual de 64 bits (pHash/aHash). Este umbral se adopta como criterio conservador, pues busca capturar transformaciones leves (re-escalado, recortes menores o cambios sutiles) sin colapsar imágenes distintas pero semánticamente similares en la misma clase.

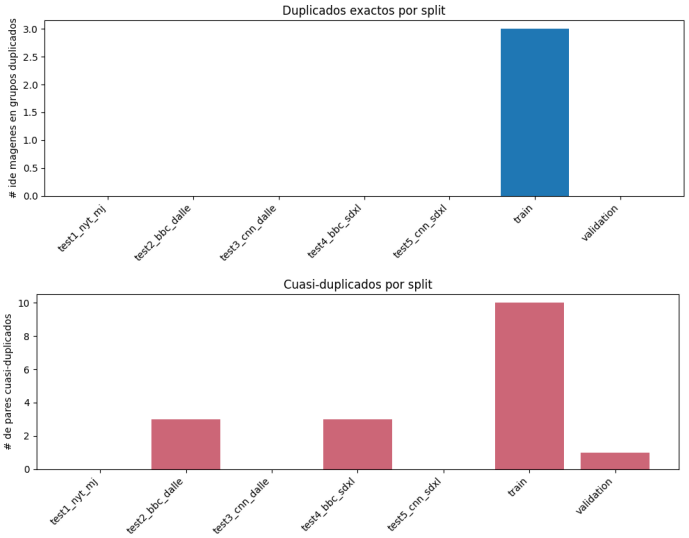


Figura 3.7: Número de duplicados y cuasi-duplicados por split.

En el split *train* se detectaron **3** duplicados exactos y **10** cuasi-duplicados; en *validation* se detectaron **0** duplicados exactos y **1** cuasi-duplicado. En los splits de prueba, no se detectaron duplicados exactos dentro de cada split (0), y se observaron **3** cuasi-duplicados en *BBC+DALL-E 3* y **3** cuasi-duplicados en el split *BBC+SDXL*.

No se encontraron duplicados exactos entre *train* y ninguno de los splits de prueba (0 colisiones exactas). Bajo el criterio perceptual, se observó un solapamiento mínimo entre *train* y *validation* (**1** colisión con ≤ 5 de Hamming), lo cual sugiere una fuga potencial marginal si se utilizara *validation* como evaluación. En contraste, se detectó un solapamiento elevado entre algunos splits de prueba fuera de dominio que comparten la misma fuente real: **209** imágenes en común entre *BBC+DALL-E 3* y *BBC+SDXL*, y **211** imágenes en común entre *CNN+DALL-E 3* y *CNN+SDXL*. Este hallazgo indica que dichos subconjuntos OOD no son independientes entre sí.

Dado que el refinamiento se realiza únicamente con *train*, la ausencia de colisiones exactas entre *train* y los splits de prueba reduce el riesgo de fuga directa hacia una eventual evaluación externa.

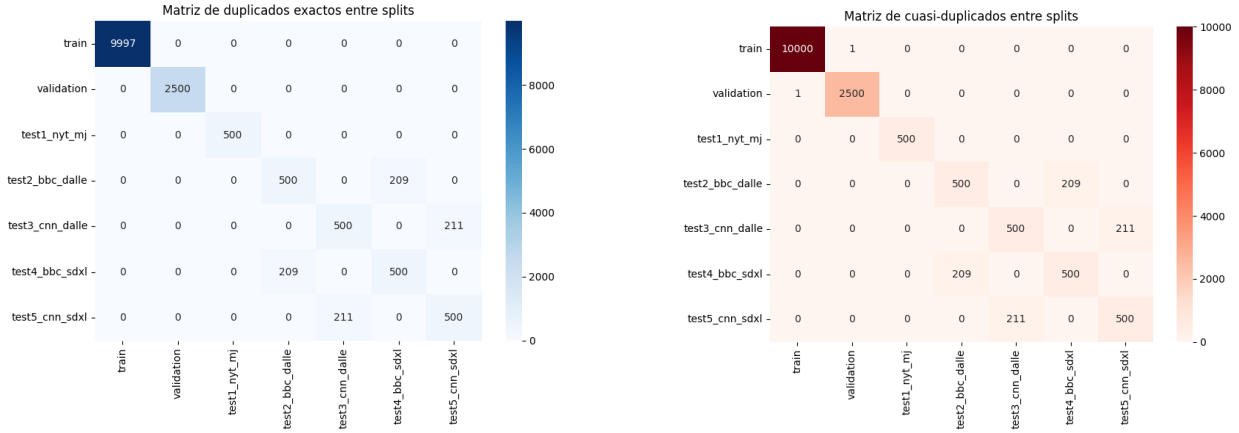


Figura 3.8: Número de colisiones entre splits.

Los duplicados internos de *train* son de orden muy bajo (3/10.000) y, aun eliminándolos, el impacto en el tamaño efectivo sería marginal; por ello, no se condicionó el protocolo de entrenamiento a una depuración adicional. Por otro lado, el solapamiento sustancial entre splits OOD de prueba sugiere que agregar esos subconjuntos sin control puede inflar métricas y sobre-representar ciertas imágenes reales; esto contribuyó a la decisión de no usar los splits *validation/test* de MiRAGeNews como evaluación principal y, en su lugar, diseñar un conjunto de validación personalizado y externo.

Adicionalmente, se analizó la variación de resolución en el split *train*. La Figura 3.9 muestra distribuciones multimodales: la mayoría de imágenes se concentra en resoluciones medias (picos alrededor de 600-700 px de ancho y 350-450 px de alto), y existe una cola hacia resoluciones altas cercanas a 2000 px en una dimensión. Esta heterogeneidad se gestiona mediante el preprocesamiento requerido por cada modelo (re-escalado a tamaño de entrada fijo), por lo que la resolución original no se utiliza como señal explícita para la clasificación.

En cuanto al formato, en este proyecto las imágenes del split *train* se extrajeron en su formato original como archivos `.png`, lo que reduce el riesgo de introducir artefactos de compresión con pérdida durante el pipeline de carga.

3.3. Aplicación de la técnica de aprendizaje por transferencia a los modelos

En esta sección se describen los procedimientos de preprocesamiento, configuración de refinamiento y especificaciones de cómputo utilizadas para adaptar los modelos seleccionados a la tarea de clasificación de imágenes noticiosas reales vs sintéticas, utilizando el conjunto de datos MiRAGeNews como base para el entrenamiento de los modelos. Dado que los modelos seleccionados tienen arquitecturas y preentrenamientos diferentes, se han aplicado técnicas de aprendizaje por transferencia específicas para cada uno, con el objetivo de maximizar su rendimiento en la tarea objetivo sin incurrir en sobreajuste o pérdida de generalización. La Figura 3.10 muestra el esquema de la

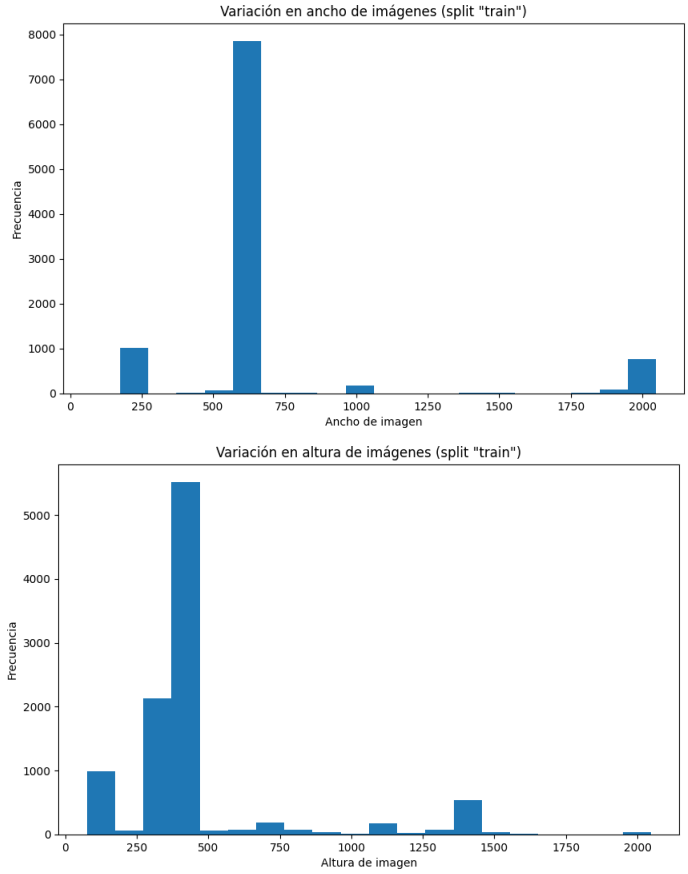


Figura 3.9: Distribución de resolución (ancho y alto) en el split *train* de MiRAGeNews.

arquitectura.

3.3.1. Preprocesamiento de imágenes para cada modelo

El preprocesamiento de las imágenes se adaptó a los requisitos específicos de cada modelo, teniendo en cuenta su arquitectura subyacente y el modo en que cada implementación interactúa con el conjunto de datos. Dado que los modelos extraídos de literatura científica tienen repositorios públicos con código de entrenamiento, se siguieron las indicaciones de preprocesamiento recomendadas por los autores para cada uno de ellos, de forma que las entradas cumplieran estrictamente los requerimientos de cada arquitectura.

Si bien el conjunto de datos se puede importar mediante las librerías de Hugging Face, los modelos extraídos de literatura científica no cuentan con esta facilidad; por ello, se implementó un preprocesamiento manual partiendo de la extracción de las imágenes del split *train* en su formato original `.png`. Posteriormente, se aplicó un preprocesamiento específico por modelo siguiendo las indicaciones de los autores en sus respectivos repositorios.

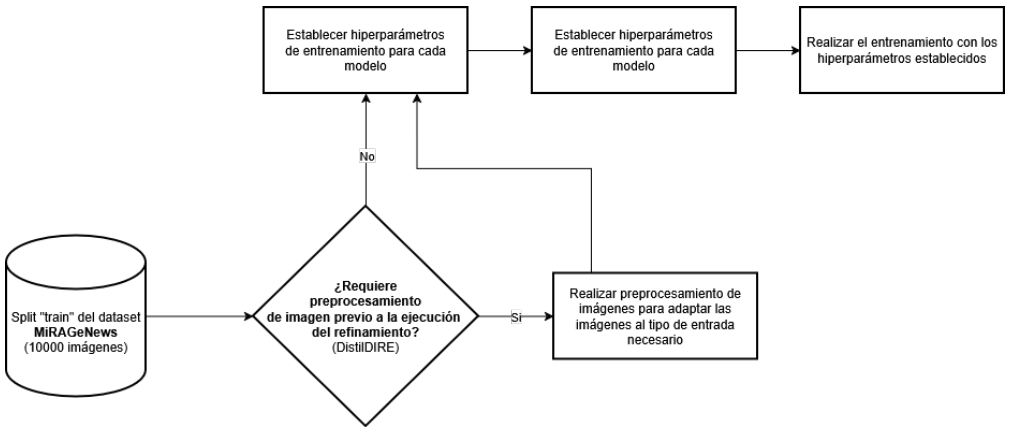


Figura 3.10: Flujo que describe las acciones tomadas para la aplicación de la técnica de aprendizaje por transferencia.

Con el fin de asegurar reproducibilidad y comparabilidad, se documenta explícitamente el protocolo de entrada común: las imágenes se trabajaron en formato `.png` y (i) se convirtieron a RGB cuando aplica, (ii) se re-escalaron a un tamaño fijo de 224×224 píxeles (tamaño final de entrada utilizado en esta comparativa), y (iii) se normalizaron según el backbone de cada modelo (p.ej., estadísticas/convenciones definidas por CLIP, ResNet o el *image processor* correspondiente). El entrenamiento se realizó con tamaño de batch 16 (Tabla 3.4). Además, el preprocesamiento utilizado en entrenamiento y validación fue idéntico para cada modelo, evitando discrepancias entre splits; en particular, no se introdujeron transformaciones adicionales ad-hoc fuera de las que define cada repositorio.

El código fuente del modelo UniversalFakeDetect [45] incluye la función de preprocesamiento que normaliza y redimensiona las imágenes de entrada según los requisitos del backbone CLIP. En esta comparativa, esta función se empleó sin modificaciones: cada imagen se convirtió a RGB y se reescaló a 224×224, aplicando la normalización estándar del backbone CLIP usado por el repositorio. De este modo, el entrenamiento se ejecutó directamente sobre las imágenes del conjunto de datos (sin transformaciones externas adicionales). En el refinamiento por transferencia se mantuvo congelado el backbone CLIP y se entrenaron únicamente los pesos de la capa de clasificación.

A diferencia de los otros modelos presentados en este proyecto, DistilDIRE requiere una transformación adicional previa al entrenamiento. Dada una imagen de entrada x_0, el proceso de inversión DDIM estima una representación latente asociada al ruido inicial del proceso generativo,

$$\varepsilon_0 = \Phi_{\mathrm{DDIM}}(x_0),$$

donde Φ_{DDIM} representa la operación de inversión sobre el modelo de difusión preentrenado. Esta transformación permite aproximar el estado de ruido desde el cual la imagen pudo haber sido generada, proporcionando una señal complementaria a la información visual contenida en la imagen original. En la práctica, para cada imagen x_0 se realizó: (i) conversión a RGB y re-escalado

a 224×224, y (ii) un proceso de inversión DDIM para estimar el ruido inicial ε_0 correspondiente. Este preprocesamiento previo al entrenamiento tomó aproximadamente **160 horas** para el split *train* completo.

Para hacer el pipeline reproducible y evitar recomputar la inversión en cada época, los resultados se persistieron en disco: cada imagen se almacenó en formato 224×224 y se acompañó de su tensor de ruido en un archivo `.pt` por imagen (con un esquema de identificación consistente entre imagen y tensor). Este procedimiento habilita la construcción del vector de entrada utilizado por el estudiante. Formalmente, el modelo opera sobre la representación conjunta

$$z = \Psi(x_0, \varepsilon_0),$$

donde $\Psi(\cdot)$ denota el mecanismo de fusión implementado por la arquitectura. A partir de esta representación, el clasificador produce una predicción binaria sobre la autenticidad de la imagen. En el refinamiento por transferencia se mantuvo congelado el backbone ResNet-50 y se entrenaron los parámetros del estudiante según el protocolo del repositorio.

Similar a UniversalFakeDetect, el código fuente del modelo D^3 [46] incluye la función de preprocesamiento para las imágenes de entrada. Esta función aplica el redimensionamiento a 224×224 y un conjunto de transformaciones de aumento definidas por los autores. Adicionalmente, para cada imagen original x se genera una segunda vista

$$\tilde{x} = T(x),$$

donde $T(\cdot)$ representa la operación de *patch shuffling*. El conjunto de entrenamiento queda entonces definido como pares

$$(x, \tilde{x}),$$

permitiendo que la arquitectura aprenda relaciones de consistencia y discrepancia entre ambas representaciones durante el refinamiento. Desde una perspectiva de implementación, cada iteración de entrenamiento ejecuta la secuencia:

$$x \to (x, \tilde{x}) \to (h_1, h_2) \to h_f \to \hat{y},$$

donde la generación de la vista distorsionada constituye una transformación determinística previa al backbone CLIP, la extracción de características se realiza mediante pesos congelados y únicamente el bloque de fusión y clasificación participa en la optimización mediante retropropagación. Gracias a esto, el entrenamiento se realizó usando las imágenes del conjunto de datos sin modificaciones adicionales fuera del repositorio. Adicionalmente, el conjunto de parámetros entrenables se restringió a

$$\theta = \{\theta_{\text{att}}, \theta_{\text{fc}}\},$$

donde θ_{att} corresponde a la capa de autoatención y θ_{fc} a la capa completamente conectada utilizada para clasificación. Los parámetros del backbone CLIP permanecieron congelados durante todo el proceso de refinamiento,

$$\theta_{\text{CLIP}} = \text{constante},$$

reduciendo la complejidad del ajuste y favoreciendo la transferencia de conocimiento desde las representaciones preentrenadas.

Cabe resaltar que tanto el conjunto de datos como los modelos extraídos de Hugging Face proveen disponibilidad mediante las librerías de importación, incluyendo el acceso directo al contenido de los splits y a los pesos preentrenados. Considerando las similitudes de integración (uso de *image processors* de `Transformers`), se aplicó un mismo patrón de preprocesamiento para los tres modelos: conversión a RGB, re-escalado a 224×224 y normalización según el *processor* del modelo. Además, en cada caso, se congelaron todas las capas del backbone y se entrenó únicamente la última capa de clasificación. El fragmento de código de 3.1 describe la función de preprocesamiento utilizada tanto para el entrenamiento como para la evaluación de los modelos de Hugging Face.

```python
def preprocess_batch(examples, processor):
    """Preprocesa un batch compatible con Hugging Face Datasets.
    - Convierte imagenes a RGB.
    - Aplica el processor del modelo (incluye re-escalado a 224x224 y
    normalizacion).
    - Devuelve tensores listos para entrenamiento/evaluaci n.
    """

    # ‘examples‘ es un diccionario con listas (p.ej., examples["image"] y examples
    ["label"]).
    images_rgb = [img.convert("RGB") for img in examples["image"]]
    inputs = processor(images=images_rgb, return_tensors="pt")
    # La salida de ‘processor‘ incluye "pixel_values" con las im genes
    preprocesadas.
    pixel_values = inputs["pixel_values"]
    labels_tensor = torch.tensor(examples["label"], dtype=torch.long)
    # Devuelve un diccionario con las im genes preprocesadas y las etiquetas como
     tensores.
    return {"pixel_values": pixel_values, "labels": labels_tensor}
```

Listing 3.1: Función de preprocesamiento aplicada a las imágenes para los modelos extraídos de Hugging Face.

3.3.2. Configuración del refinamiento para cada modelo

Con el objetivo de mantener un protocolo homogéneo y comparable entre arquitecturas, el refinamiento se formuló como un problema de optimización supervisada. Sea

$$\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N}$$

el conjunto de entrenamiento, donde x_i representa una imagen y $y_i \in \{0, 1\}$ su etiqueta de autenticidad. El objetivo consiste en encontrar los parámetros entrenables θ que minimicen una función de pérdida

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(f_\theta(x_i), y_i\right),$$

donde $\mathcal{L}$ corresponde a la función de pérdida utilizada por cada arquitectura (entropía cruzada, pérdida sigmoidal o formulaciones híbridas).

Bajo este esquema, se fijaron hiperparámetros comunes (número de épocas y tamaño de batch) y únicamente se ajustaron aquellos parámetros necesarios para garantizar estabilidad numérica y compatibilidad con cada implementación.

En todos los casos, el refinamiento se planteó como aprendizaje por transferencia ligero. Sea

$$\theta = \theta_f \cup \theta_c,$$

donde θ_f representa los parámetros del backbone preentrenado y θ_c los parámetros de clasificación específicos de cada modelo.

La estrategia adoptada consistió en mantener

$$\nabla_{\theta_f} \mathcal{L} = 0,$$

para todos los modelos evaluados, mientras que únicamente los parámetros de decisión satisfacen

$$\theta_c \leftarrow \theta_c - \eta \nabla_{\theta_c} \mathcal{L}.$$

De esta forma, el refinamiento se restringe a las capas finales y módulos específicos definidos por cada arquitectura (cabezas lineales, bloques de atención o componentes equivalentes), preservando el conocimiento adquirido durante el preentrenamiento y reduciendo el riesgo de sobreajuste sobre MiRAGeNews.

Las tasas de aprendizaje de la Tabla 3.4 se seleccionaron dentro de rangos típicos de refinamiento cuando se entrenan únicamente capas finales sobre un backbone congelado (órdenes de magnitud 10^{-5}–10^{-4}, con una excepción de 5×10^{-4}), buscando un compromiso conservador entre (i) evitar divergencia/instabilidad y (ii) permitir que la cabeza alcanzara un ajuste útil dentro del presupuesto fijo de épocas. Dado que no se empleó scheduler, la tasa se mantuvo constante durante el refinamiento.

Dado que los repositorios de los modelos extraídos de literatura científica sugieren el uso de tasas de aprendizaje en el rango 10^{-4} para sus respectivas arquitecturas, se optó por mantener esa referencia como punto de partida para la mayoría de los modelos, realizando ajustes a este valor con respecto a los modelos de Hugging Face para asegurar estabilidad numérica (En el caso de 5×10^{-5} para Smogy y 2×10^{-5} para prithivMLmods, se debe a que demostraron poseer mejor convergencia en pruebas preliminares).

Se fijó un valor de **5 épocas** para todos los modelos por dos motivos: (i) garantizar comparabilidad bajo el mismo presupuesto de cómputo, y (ii) controlar el costo experimental, especialmente en modelos con preprocesamientos costosos (como DistilDIRE, el cuál requiere inversión DDIM previa, tomando aproximadamente 160 horas para el split *train*). Bajo el esquema de backbone

congelado, el ajuste se concentra en pocas capas y, en general, converge en pocas pasadas sobre los datos; por ello, el objetivo no fue maximizar rendimiento mediante búsqueda extensiva de épocas, sino obtener un refinamiento uniforme y reproducible.

La tabla 3.4 resume la configuración de refinamiento aplicada a cada modelo.

Modelo	Tasa de Aprendizaje	Épocas	Tamaño de Batch	Optimizador	Función de Pérdida
UniversalFakeDetect	1e-4	5	16	Adam	Entropía cruzada
DistilDIRE	1e-4	5	16	Adam	Híbrido (entropía cruzada + error cuadrático medio)
D^3	1e-4	5	16	AdamW	Entropía cruzada
Smogy/SMOGY-Ai-images-detector	5e-5	5	16	AdamW	Entropía cruzada
Ateeqq/ai-vs-human-image-detector	5e-4	5	16	AdamW	Sigmoidal
prithivMLmods/deepfake-detector-model-v1	2e-5	5	16	AdamW	Sigmoidal

Cuadro 3.4: Configuración de refinamiento aplicada a cada modelo durante el proceso de entrenamiento.

En el caso particular de DistilDIRE, el entrenamiento utiliza una función objetivo híbrida compuesta por un término de clasificación y un término de destilación. De manera general, la función de pérdida puede expresarse como

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{MSE},$$

donde $\mathcal{L}_{CE}$ corresponde a la pérdida de entropía cruzada empleada para la clasificación binaria y $\mathcal{L}_{MSE}$ representa el error cuadrático medio utilizado para aproximar el comportamiento del modelo maestro. Los coeficientes λ_1 y λ_2 controlan la contribución relativa de cada término durante el refinamiento.

Es importante aclarar que no se utilizó un conjunto de validación para validar las métricas durante el entrenamiento ni para seleccionar hiperparámetros, dado que el protocolo de este proyecto se basa en un esquema de evaluación externa con un conjunto de validación personalizado. Por ello, el refinamiento se realizó exclusivamente sobre el split *train* de MiRAGeNews, y las métricas reportadas al finalizar las 5 épocas se basan en el desempeño sobre dicho split, sin realizar ajustes basados en validación.

Durante la ejecución del refinamiento se almacenó un checkpoint al finalizar cada época. Sea

$$\mathcal{M}_e = (L_e, F1_e)$$

el par de métricas asociado a la época e, donde L_e representa la pérdida observada y $F1_e$ el puntaje F1 correspondiente. El checkpoint seleccionado para evaluación posterior fue aquel que maximizó el desempeño clasificatorio y minimizó simultáneamente la pérdida de entrenamiento dentro del conjunto de modelos generados durante las cinco épocas. Al finalizar las 5 épocas, se seleccionó el checkpoint con menor pérdida y mayor puntaje F1 reportado para cada modelo. Este proceso permitió asegurar que cada modelo se entrenara bajo condiciones óptimas dentro del presupuesto definido, y que los resultados reportados reflejaran el mejor desempeño alcanzado durante el refinamiento. La Figura 3.11 muestra las curvas de pérdida de entrenamiento por época para cada modelo, evidenciando la convergencia y estabilidad del proceso de refinamiento.

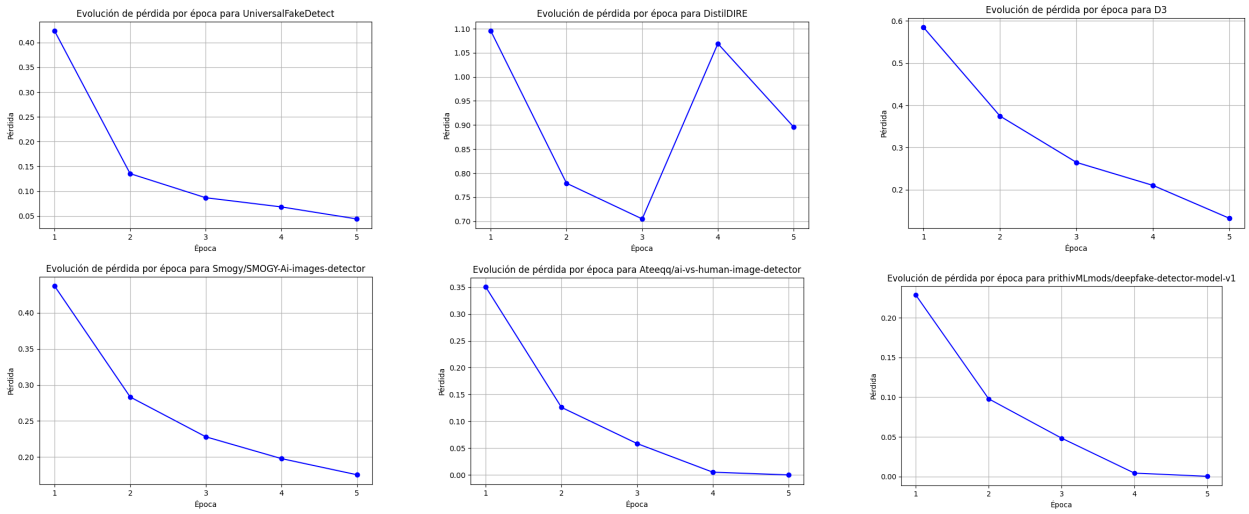


Figura 3.11: Curvas de pérdida de entrenamiento por época para cada modelo durante el proceso de refinamiento.

3.3.3. Especificaciones de cómputo utilizadas para el entrenamiento

El entrenamiento de los modelos se llevó a cabo en una infraestructura de cómputo local. Con el objetivo de mejorar la reproducibilidad y clarificar las restricciones computacionales, a continuación se detallan las especificaciones de hardware, el sistema operativo y el backend usado para ejecutar PyTorch sobre la GPU.

- **Hardware:**

Componente	Especificación
CPU	AMD Ryzen 5 5600G @ 3.9GHz (6 núcleos, 12 hilos)
GPU	AMD Radeon RX 9060 XT (16GB VRAM)
RAM	32GB DDR4 @ 3200MHz
Almacenamiento	SSD NVMe de 500GB

Cuadro 3.5: Especificaciones de hardware utilizadas para el entrenamiento de los modelos.

- **Software y plataforma:** El entorno se configuró en **Windows 10 (versión 22H2)** con **Python 3.12.5**. Para la implementación y el refinamiento de los modelos se utilizó **PyTorch 2.10** y las librerías **Transformers** y **Datasets** de Hugging Face para la gestión de modelos preentrenados y del conjunto de datos.
- **Aceleración en GPU (AMD):** El hardware emplea una GPU AMD, con versión de driver AMD **26.3.1**, el cuál permite la ejecución de PyTorch utilizando aceleración GPU en sistemas

operativos Windows. Gracias a esto, la ejecución de entrenamiento se realizó mediante **ROCm v7.2.1** (backend *HIP/ROCm*).

Para cuantificar el costo computacional del refinamiento, la Tabla 3.6 reporta los tiempos totales de entrenamiento registrados para cada modelo bajo el protocolo descrito en esta sección.

Modelo	**Tiempo de entrenamiento**
UniversalFakeDetect	2 h 47 min 18 s
DistilDIRE	2 h 31 min 56 s
D^3	3 h 05 min 42 s
Smogy/SMOGY-Ai-images-detector	1 h 10 min 50 s
Ateeqq/ai-vs-human-image-detector	1 h 08 min 35 s
prithivMLmods/deepfake-detector-model-v1	1 h 09 min 15 s

Cuadro 3.6: Tiempos de entrenamiento (refinamiento) registrados para cada modelo en el hardware descrito.

3.4. Diseño del conjunto de datos de validación

Si bien el conjunto de datos MiRAGeNews ya incluye una partición de validación y varias particiones de prueba, estas poseen debilidades identificadas en la sección de análisis exploratorio. En particular, la presencia de duplicados y cuasi-duplicados entre algunos splits de prueba fuera de dominio puede inflar artificialmente las métricas de rendimiento. Adicionalmente, el uso de imágenes sintéticas generadas con modelos de 2023 puede no reflejar la calidad y diversidad del ecosistema actual de generación de imágenes.

Con el fin de complementar la evaluación y abordar las limitaciones anteriores, se propone un conjunto de evaluación externo y personalizado. Este conjunto busca (i) incorporar imágenes sintéticas generadas con modelos contemporáneos y (ii) mantener un balance de clases que haga las métricas fácilmente interpretables. Sin embargo, dado su tamaño ($n = 100$), este conjunto se presenta explícitamente como una evaluación externa exploratoria (no definitiva), útil para analizar robustez y tendencias, pero insuficiente por sí sola para sostener conclusiones fuertes.

La creación del conjunto de evaluación personalizado se divide en dos partes: la construcción del split de imágenes sintéticas ("fake") y la construcción del split de imágenes reales ("real"), cada una con consideraciones metodológicas orientadas a la calidad, diversidad y trazabilidad del proceso. La Figura 3.12 muestra un esquema general del diseño.

Para garantizar el balance en la distribución de clases, se establece que el conjunto contenga 50 imágenes reales y 50 imágenes sintéticas, totalizando 100 imágenes. Con el objetivo de reducir "atajos"basados en metadatos o contenedores de archivo, las imágenes reales y sintéticas se homogenizaron en el mismo formato de archivo (PNG). No se aplicaron etapas adicionales de post-procesamiento (recorte, compresión, normalización o edición) sobre las imágenes, más allá de la homogenización de formato.

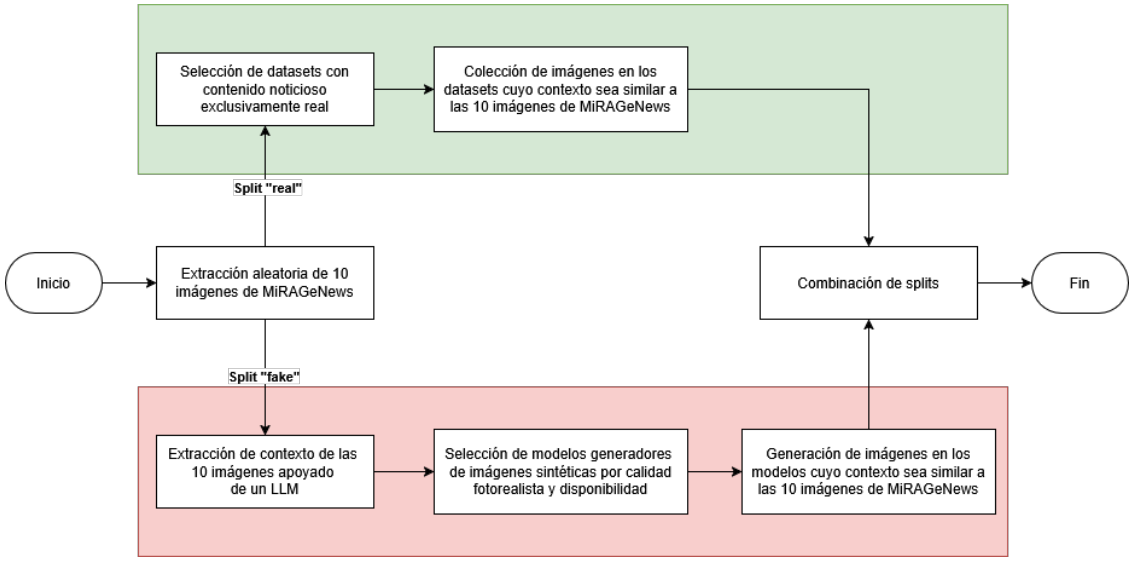


Figura 3.12: Esquema general del diseño del conjunto de datos de validación personalizado.

3.4.1. Creación del split de imágenes sintéticas ("fake")

Para mantener un criterio metodológico riguroso y replicable, se definió el split "fake.ª partir de 10 imágenes sintéticas seleccionadas aleatoriamente desde MiRAGeNews. A partir de estas 10 imágenes, se realizó un paso de extracción de contexto con apoyo de un LLM, con el objetivo de convertir cada imagen en una descripción breve y reutilizable como prompt de generación en distintos modelos text-to-image.

El uso del LLM para esta etapa se justifica por dos razones: (i) estandariza la estructura y longitud de los prompts, y (ii) permite documentar y reproducir el proceso de generación a partir de un procedimiento uniforme. No obstante, este enfoque también puede introducir sesgo, particularmente en términos de homogeneidad semántica en las descripciones o sesgo del propio LLM, por lo que se asume como una limitación metodológica y se documenta con transparencia.

Cabe resaltar que la descripción de cada imagen debe mantenerse concisa, debido a que los modelos generativos de imágenes suelen tener limitaciones estrictas en la longitud efectiva del prompt (y prompts demasiado largos tienden a introducir detalles no controlados). Por ende, se diseñó el siguiente prompt para producir únicamente descripciones breves que posteriormente sirvieran como prompts de generación:

"Describe los contenidos de esta imagen en un parágrafo de 30 palabras o menos, detallando concisamente los hechos que se están retratando"

El modelo LLM utilizado para la extracción de contexto fue GPT-5.3-Instant, accedido en la interfaz web ChatGPT de OpenAI durante los días 16-18 de marzo de 2026, y con parámetros de generación predeterminados (temperatura 1.0, sin reintentos). Se documentó la conversación completa utilizada para esta extracción, incluyendo los prompts y respuestas, en el anexo 6.1.

Después de obtener las descripciones de cada una de las 10 imágenes, se seleccionaron modelos generativos de imágenes representativos del ecosistema contemporáneo de generación de imágenes sintéticas, con el fin de generar nuevas imágenes a partir de los prompts extraídos. Para mantener criterios de selección trazables, se consultaron fuentes de evaluación y benchmarking para modelos generativos; específicamente, la plataforma Arena AI [12], la cual ofrece leaderboards de desempeño en tareas texto-a-imágen. La Figura 3.13 muestra un fragmento del leaderboard consultado.

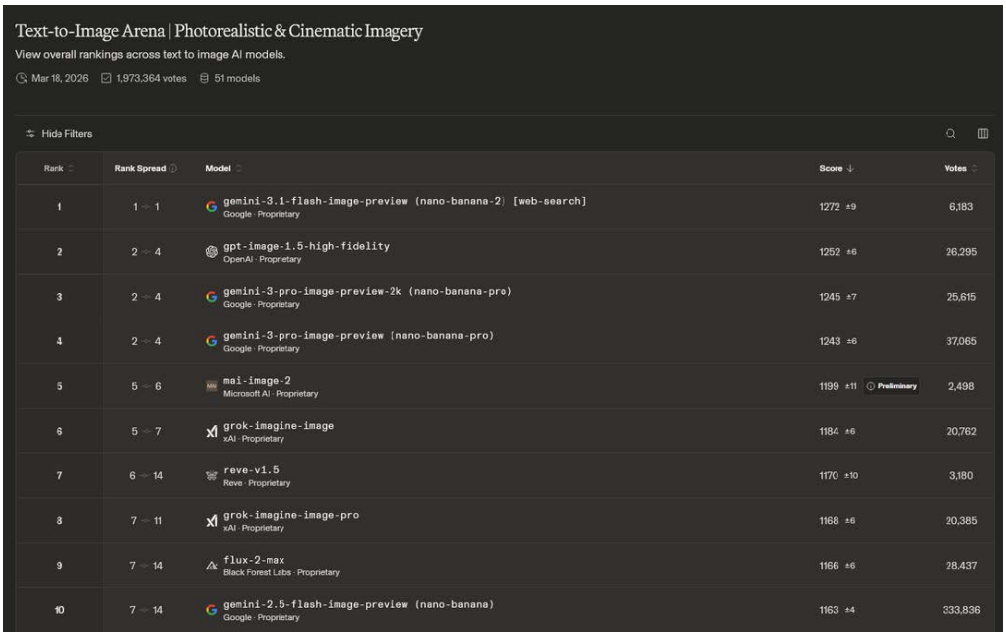


Figura 3.13: Fragmento del leaderboard para modelos texto-a-imágen en Arena AI, incluyendo los puntajes obtenidos por diferentes modelos generativos de imágenes en esta tarea (Fuente: [12]).

Para entender el puntaje y así mismo seleccionar los modelos generativos de imágenes a utilizar, es necesario entender la métrica de evaluación utilizada en el leaderboard de Arena AI. En el caso de esta plataforma, el sistema de ranking funciona con puntuaciones Elo, en las que los modelos comienzan con una puntuación de referencia de entre 1000 y 1200, y los puntos se intercambian en función de los votos de los usuarios, con cambios más grandes cuando un modelo de menor Elo se vota en preferencia sobre modelos con mayor Elo. La obtención y substracción de puntos se realiza automáticamente según los resultados de los votos realizados por usuarios que usan esta plataforma, tras enviar prompts en el modo "Arena" y recibir resultados anónimos de dos modelos dependiendo de la tarea realizada, en el cuál pueden votar por cuál modelo consideran que tuvo el mejor resultado. Además, las puntuaciones Elo de la plataforma incluyen intervalos de confianza, donde se muestra una "posición bruta" basada en estimaciones puntuales de cada modelo, además de un "rango de variación" que indica las mejores y peores posiciones posibles teniendo en cuenta el rango de calidad de los resultados del modelo [47].

Si bien este sistema de puntajes no es una métrica de evaluación objetiva, sino que refleja las

preferencias subjetivas de los usuarios que participan en la plataforma, continúa siendo un indicador de la calidad de los modelos generativos disponibles actualmente. Dado que se requiere seleccionar modelos generativos de imágenes que sean tanto accesibles como representativos del ecosistema actual, se tuvieron en cuenta aquellos modelos que fueran de acceso gratuito o con una versión gratuita disponible, que estuvieran en el top 50 del leaderboard de generación de imágenes a partir de texto en Arena AI, y para tener diversidad en la selección con el fin de prevenir la homogenización de resultados, se tiene en cuenta como máximo un modelo por proveedor de modelos generativos de imágenes presentes en el leaderboard. Siguiendo estos criterios, se seleccionaron los siguientes modelos generativos de imágenes para la generación de las imágenes sintéticas del split "fake"del conjunto de datos de validación personalizado:

- `gemini-2.5-flash-image` (Elo: 1163, Intervalo de confianza: ±4, Proveedor: Google)
- `qwen-image-2512` (Elo: 1148, Intervalo de confianza: ±6, Proveedor: Alibaba)
- `mai-image-1` (Elo: 1102, Intervalo de confianza: ±5, Proveedor: Microsoft AI)
- `flux-2-klein-9b` (Elo: 1086, Intervalo de confianza: ±6, Proveedor: Black Forest Labs)
- `glm-image` (Elo: 999, Intervalo de confianza: ±13, Proveedor: Z.ai)

En total se generaron 50 imágenes sintéticas, distribuidas de manera uniforme entre los cinco generadores seleccionados (10 imágenes por generador). Como parámetros de generación, se utilizó el valor de semilla 1111 para garantizar la reproducibilidad de los resultados, y se añadió variantes de relación de aspecto (1:1, 4:3, 16:9) para cada imagen, con el objetivo de aumentar la diversidad visual del conjunto.

Debido a la naturaleza del contenido de algunas de las imágenes sintéticas previamente extraídas del dataset MiRAGeNews, se tuvieron que realizar ajustes en las descripciones obtenidas para que estas pudieran ser utilizadas como prompts en los modelos generativos de imágenes seleccionados. Esto se debe a que algunas de las imágenes sintéticas extraídas del dataset MiRAGeNews retratan escenas de violencia o contenido explícito, lo cuál no es permitido por algunos de los modelos generativos de imágenes seleccionados.

Para abordar esta limitación, se realizaron ajustes pequeños en las descripciones obtenidas para eliminar cualquier referencia a contenido explícito o violento, manteniendo la esencia de la escena retratada pero sin incluir detalles que pudieran bloquear la generación de la imagen de parte de los modelos generativos de imágenes. El criterio principal para realizar estos ajustes fue el siguiente: Para cada prompt que haya sido anulado debido a su descripción explícita por los modelos, se realizó la modificación de las palabras que se consideran más aludidas a contenido explícito o violento, y una vez se modificaron, se realizó un reintento de generación.

3.4.2. Creación del split de imágenes reales ("real")

Para la creación del split "real", se optó por una estrategia de selección manual a partir de un conjunto de datos especializado en imágenes de noticias reales, con el objetivo de garantizar

autenticidad y relevancia contextual. Para esto, se utilizó el dataset VisualNews [48], el cual contiene imágenes extraídas de artículos de noticias reales y ha sido utilizado en investigaciones previas relacionadas con análisis y automatización de etiquetado para contenido noticioso.

La selección de las imágenes reales se realizó de manera manual en el split "bbc"de VisualNews (imágenes originarias de BBC News). Para cada contexto de las 10 imágenes obtenidas en el procedimiento de extracción aleatoria del split "real" de MiRAGeNews, se seleccionaron manualmente 5 imágenes de VisualNews por similitud en términos de contenido y contexto, totalizando 50 imágenes reales. Dado que una imagen puede variar sin cambiar su contenido principal (iluminación, ángulo, encuadre), el criterio de selección priorizó el contexto principal retratado, buscando coherencia temática entre clases y una comparación más justa en la tarea de clasificación real vs. sintética. La selección manual, aunque permite control cualitativo del contexto, introduce un riesgo de sesgo, como escoger inadvertidamente imágenes "más fáciles.º con señales visuales particulares. La selección se realizó con la participación del equipo de investigación, utilizando el siguiente protocolo:

1. Para cada una de las 10 imágenes obtenidas aleatoriamente, se describió el contexto principal de la imágen (hechos retratados, personas que aparecen, ambiente en el que se tomó la imágen).

2. Del split "bbc"de VisualNews, Se seleccionaron las 5 imágenes que tuvieran tanto mayor similitud fotográfica a la imágen original, como mayor coincidencia con el contexto descrito anteriormente.

Una limitación relevante de esta etapa es la antigüedad de las imágenes en VisualNews: la mayoría proviene de artículos publicados entre 2010 y 2015. Esto puede introducir un *confound* importante frente a imágenes sintéticas contemporáneas, ya que los modelos podrían explotar señales de antigüedad (estilo editorial, compresión, resolución típica, estética fotográfica) en lugar de "autenticidad". Por lo anterior, los resultados en este conjunto deben interpretarse con cautela y enmarcarse como evidencia exploratoria de robustez. Finalmente, al igual que en el split "fake", se mantuvo la homogenización de formato de archivo para reducir atajos basados en el tipo de archivo.

Capítulo 4

Resultados

Después de haber aplicado la técnica de refinamiento a los modelos de detección seleccionados y haberlos evaluado sobre el conjunto de datos personalizados, se presentarán los resultados obtenidos en esta sección. Se describirán las métricas de desempeño que alcanzaron los modelos sobre el conjunto de datos personalizado, así como una discusión sobre por qué aquellos modelos obtuvieron ciertos resultados.

En todos los casos, la evaluación se realizó sobre el conjunto de datos de validación personalizado descrito en la Figura 3.12, compuesto por 100 imágenes balanceadas (50 reales y 50 sintéticas). Para el cálculo de precisión, recall, puntaje F1 y ROC-AUC se define la clase positiva como la clase "fake" (imagen sintética, etiqueta 1). Las matrices de confusión reportan el desempeño al umbral de decisión utilizado durante la inferencia, mientras que las curvas ROC y su área bajo la curva (ROC-AUC) sintetizan la separabilidad del modelo a través de múltiples umbrales.

4.1. Validación de resultados para el modelo UniversalFakeDetect

La Figura 4.1 muestra la matriz de confusión y la curva ROC obtenidas para UniversalFakeDetect (UFD). El modelo alcanzó una exactitud de 0,9100, con precisión de 0,9020, recall de 0,9200, puntaje F1 de 0,9109 y un ROC-AUC de 0,9740. En esta comparativa, UFD se adaptó mediante aprendizaje por transferencia ligero, manteniendo congelado el backbone CLIP y entrenando únicamente la cabeza de clasificación (sin *fine-tuning* del backbone).

En la matriz de confusión se observa un comportamiento relativamente balanceado entre errores de tipo falso positivo y falso negativo: 5 imágenes reales fueron clasificadas erróneamente como sintéticas (falsos positivos), mientras que 4 imágenes sintéticas fueron clasificadas como reales (falsos negativos). En términos prácticos, esto sugiere que, bajo el umbral utilizado, UFD captura gran parte del contenido sintético en esta muestra, manteniendo a la vez un margen acotado de "alarmas falsas" sobre imágenes auténticas.

Esta estabilidad es coherente con el diseño metodológico descrito en el Capítulo 3: UFD opera sobre un espacio de características de CLIP congelado y ajusta únicamente un clasificador superficial. Una interpretación plausible (planteada aquí como hipótesis, no como causalidad demostrada) es que, al no ajustar el backbone en un dataset acotado, el sistema reduce la probabilidad de especializarse en artefactos idiosincráticos del conjunto de entrenamiento (p. ej., patrones particulares presentes en MiRAGeNews) y podría conservar señales más generales que se transfieren mejor a imágenes sintéticas producidas por generadores contemporáneos (`gemini-2.5-flash-image`, `qwen-image-2512`, `mai-image-1`, `flux-2-klein-9b` y `glm-image`). En ese marco, el ROC-AUC al-

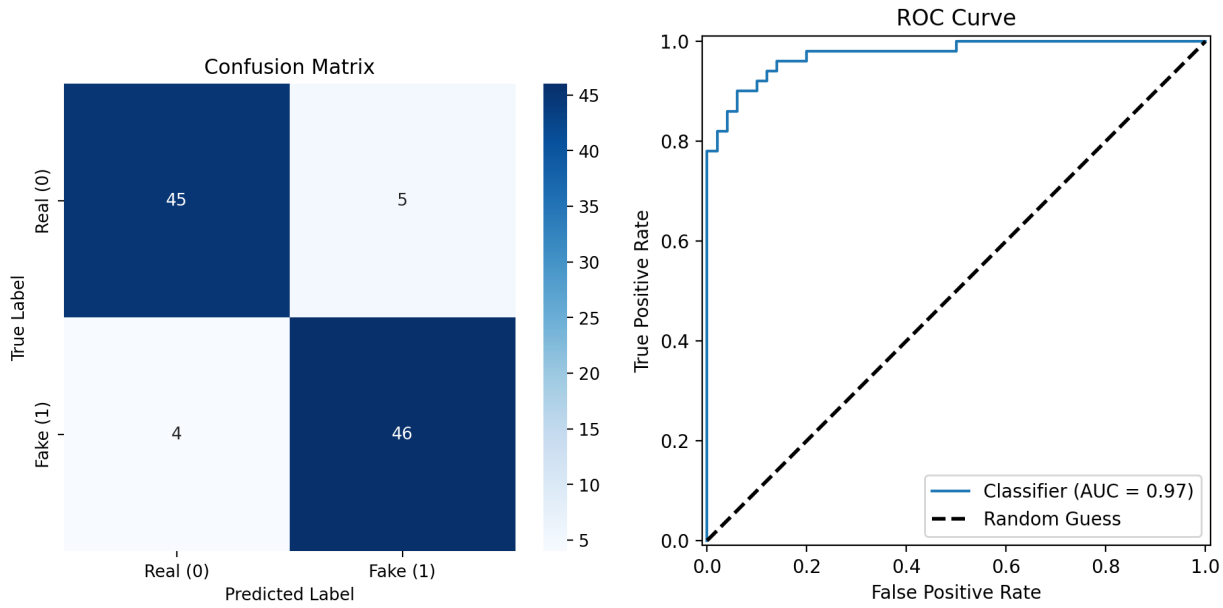


Figura 4.1: Matriz de confusión (izquierda) y curva ROC (derecha) para UniversalFakeDetect en el conjunto de validación personalizado.

to sugiere que el modelo asigna puntuaciones sistemáticamente mayores a la clase sintética a través de múltiples umbrales.

El umbral de decisión utilizado en la inferencia para UFD fue el valor por defecto (0.5), y dado que el ROC-AUC sintetiza separabilidad por umbral, un valor de 0,9740 indica que existe margen para ajustar el umbral de decisión según el objetivo operativo. Por ejemplo, en un contexto periodístico podría priorizarse (i) recall sobre "fake" para minimizar falsos negativos (dejar pasar imágenes sintéticas), o (ii) precisión para minimizar falsos positivos (etiquetar como sintética una imagen real) y reducir el costo reputacional y la sobrecarga de verificación manual. En la tabla 4.1 se propone un formato para documentar los ejemplos de falsos positivos y falsos negativos cometidos por UFD, lo cual puede ser útil para análisis cualitativos posteriores sobre las características de las imágenes que generan confusión en el modelo.

Los falsos positivos pueden interpretarse como un sesgo hacia señales estilísticas o de "acabado" visual: dado que las imágenes reales provienen de VisualNews (con fotografías predominantemente entre 2010 y 2015), es plausible que factores como compresión, baja resolución efectiva, ruido de sensor o posprocesado editorial (p. ej., nitidez/contraste) introduzcan patrones que el modelo confunda con irregularidades propias de la síntesis moderna. Por su parte, los falsos negativos sugieren la presencia de un subconjunto de imágenes sintéticas cuya estética y estructura local se aproximan fuertemente a la fotografía de prensa, reduciendo la utilidad de señales forenses de bajo nivel y forzando al clasificador a depender de cues más globales. En conjunto, UFD muestra un perfil apto para escenarios de filtrado inicial de desinformación visual, con la advertencia de que

Tipo	Imagen (ID/archivo)
Falso positivo	
Falso negativo	

Cuadro 4.1: Ejemplos de falso positivo (FP) y falso negativo (FN) cometidos por UFD en el conjunto de validación personalizado.

sus alarmas sobre imágenes reales deberían tratarse como "casos a revisar" en un flujo humano-en-el-bucle.

4.2. Validación de resultados para el modelo DistilDIRE

La Figura 4.2 presenta el desempeño de DistilDIRE sobre el conjunto de validación personalizado. A diferencia de UFD y D^3, DistilDIRE obtuvo una exactitud de 0,5400, con precisión de 0,5556, recall de 0,4000, puntaje F1 de 0,4651 y un ROC-AUC de 0,5680. En términos prácticos, bajo esta configuración y dominio, el modelo mostró una separabilidad débil entre imágenes reales y sintéticas.

La matriz de confusión exhibe dos problemas simultáneos: un número alto de falsos negativos (30 de 50 imágenes sintéticas clasificadas como reales) y un número considerable de falsos positivos (16 de 50 imágenes reales clasificadas como sintéticas). Este patrón sugiere que el puntaje producido por el modelo no separa de forma consistente ambas clases, lo cual se confirma por el ROC-AUC de 0,5680, muy próximo al 0,5 esperado por un clasificador no informativo.

Este resultado es consistente con la hipótesis de "desacople" entre la señal que DistilDIRE explota y la distribución visual del conjunto de validación personalizado, pero debe interpretarse con cautela: DistilDIRE depende de un preprocesamiento no trivial basado en inversión DDIM para extraer un término de ruido (ε_0) asociado a un modelo de difusión preentrenado, y luego clasifica a partir de la concatenación entre imagen y ruido. Esta estrategia suele ser más informativa cuando la distribución de datos y el mecanismo de generación sintética están alineados con el *prior* del difusor empleado en la inversión. En un escenario con (i) generadores heterogéneos de última

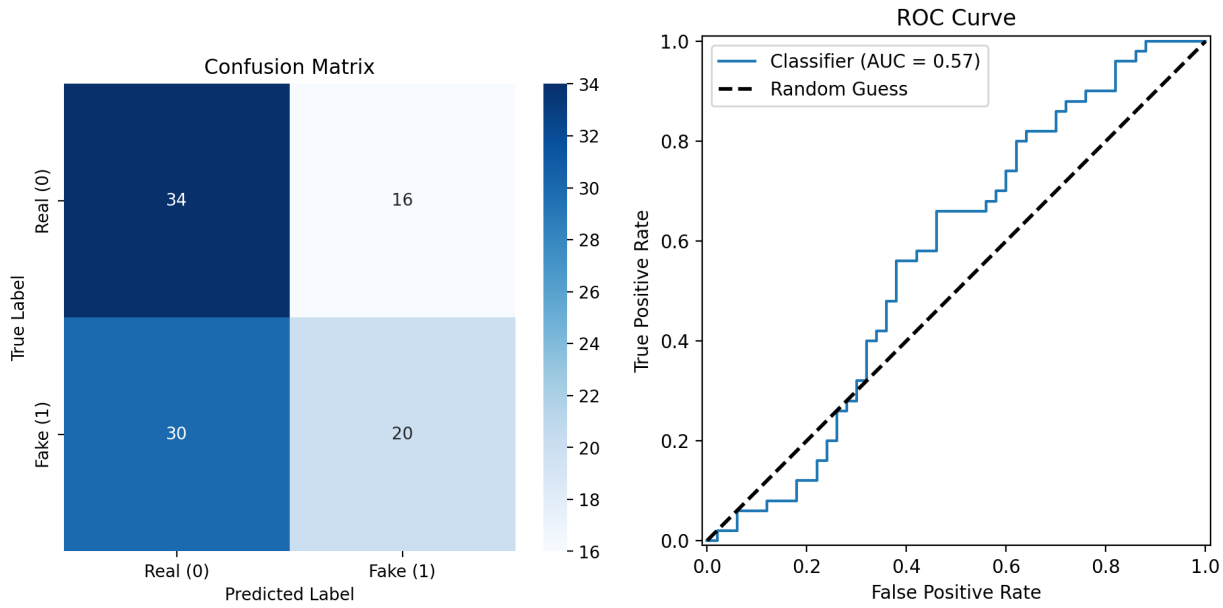


Figura 4.2: Matriz de confusión (izquierda) y curva ROC (derecha) para DistilDIRE en el conjunto de validación personalizado.

generación, potencialmente con *pipelines* internos distintos, y (ii) fotografías reales con compresión y características históricas propias del dominio periodístico, una hipótesis plausible es que el "ruido" recuperado capture variaciones que no estén asociadas de manera estable con autenticidad. De ser así, el método podría confundir degradaciones naturales (compresión, ruido, reescalados) con señales de síntesis (incrementando falsos positivos) y, a la vez, podría no reflejar trazas discriminativas en imágenes sintéticas altamente plausibles (incrementando falsos negativos).

Desde el punto de vista metodológico, antes de atribuir el bajo ROC-AUC exclusivamente a cambio de dominio, es necesario descartar que exista un fallo de implementación o de preprocesamiento. En este proyecto, el pipeline se implementó siguiendo el protocolo descrito en la metodología: conversión a RGB y re-escalado a 224×224, estimación de ε_0 por inversión DDIM, y persistencia en disco de un archivo `.pt` por imagen para garantizar consistencia entre entrenamiento y validación (sin recomputar la inversión durante el refinamiento). Aun así, la sensibilidad del método a la elección del difusor y a los hiperparámetros de inversión hace indispensable documentar y auditar explícitamente esa configuración.

Cabe aclarar que el modelo de difusión utilizado para la inversión DDIM fue el checkpoint preentrenado "256x256_diffusion_uncond", el cuál tiene una arquitectura basada en UNet y un scheduler de ruido lineal. La inversión se realizó con 50 pasos, sin guidance, y normalización estándar de imágenes. Esta configuración se eligió por ser la recomendada en la documentación original de DistilDIRE y por su compatibilidad con el tamaño de imagen utilizado (224×224). Sin embargo, dada la heterogeneidad del dominio noticioso y la variedad de generadores presentes, es plausible

que esta elección no capture adecuadamente las características relevantes para distinguir imágenes reales de sintéticas, lo que podría explicar el bajo desempeño observado.

La implicación principal es metodológica y práctica: bajo la configuración evaluada, DistilDIRE no proporciona una frontera de decisión operativamente confiable para el dominio noticioso. No obstante, este hallazgo no debe interpretarse como evidencia de que DistilDIRE sea "malo" en general, ya que la literatura reporta un desempeño alto en benchmarks específicos de difusión (p.ej., DiffusionForensics) bajo configuraciones controladas [8]. En cambio, los resultados aquí son compatibles con la idea de que el enfoque es sensible a la alineación entre el difusor usado para extraer ε_0 y los generadores presentes, y la distribución visual del dominio (compresión, degradaciones y características históricas). Por ello, más que ajustar sólo el umbral (en este caso, el umbral por defecto de 0.5), las vías razonables de mejora incluyen recalibrar la extracción de ruido con un difusor más alineado al dominio, aumentar la capacidad de adaptación durante el refinamiento (evitando congelar componentes críticos), o reentrenar con muestras sintéticas y reales que representen mejor el ecosistema contemporáneo de imágenes.

4.3. Validación de resultados para el modelo D^3

La Figura 4.3 resume el desempeño del detector D^3 (*Discrepancy Deepfake Detector*). El modelo obtuvo una exactitud de 0,9000, con precisión de 0,9167, recall de 0,8800, puntaje F1 de 0,8980 y ROC-AUC de 0,9732. Al igual que UniversalFakeDetect (UFD), estos valores reflejan una separabilidad alta en el conjunto de validación personalizado; sin embargo, dadas (i) las diferencias numéricas pequeñas entre ambos modelos y (ii) el tamaño acotado de la evaluación (100 imágenes), no es apropiado interpretar esta comparación como una jerarquía fuerte sin un análisis estadístico específico. En esta sección se enfatiza, por tanto, el *perfil de error* y la viabilidad operativa.

En particular, en un conjunto de 100 imágenes una diferencia de 0,0100 en exactitud equivale a un único acierto/error adicional, y el ROC-AUC de D^3 es prácticamente indistinguible del de UFD (0,9732 vs 0,9740). Por ello, el análisis se centra en cómo se distribuyen los errores (falsos positivos vs falsos negativos) y cómo podría ajustarse el umbral según el costo operativo de cada tipo de error.

En términos de errores, D^3 presenta 4 falsos positivos y 6 falsos negativos al umbral de decisión por defecto (0.5), utilizado durante la inferencia. En comparación con UFD (5 falsos positivos y 4 falsos negativos), la diferencia es leve y se manifiesta como un desplazamiento del compromiso precisión/recall: D^3 reduce marginalmente falsas alarmas sobre imágenes reales, pero omite una fracción algo mayor de imágenes sintéticas.

Un punto importante es evaluar si este patrón responde principalmente al *umbral* (calibración) más que a una diferencia real de separabilidad. Esto es consistente con el ROC-AUC alto (0,9732), que sugiere que el *ranking* de puntuaciones contiene información discriminativa robusta: al desplazar el umbral podría incrementarse el recall sobre la clase sintética a costa de aumentar falsos positivos, o viceversa, dependiendo del objetivo operativo.

Una explicación plausible surge de la propia arquitectura: D^3 integra una rama paralela con una versión corrompida de la imagen (por *patch shuffling*) y aprende artefactos mediante la discrepancia

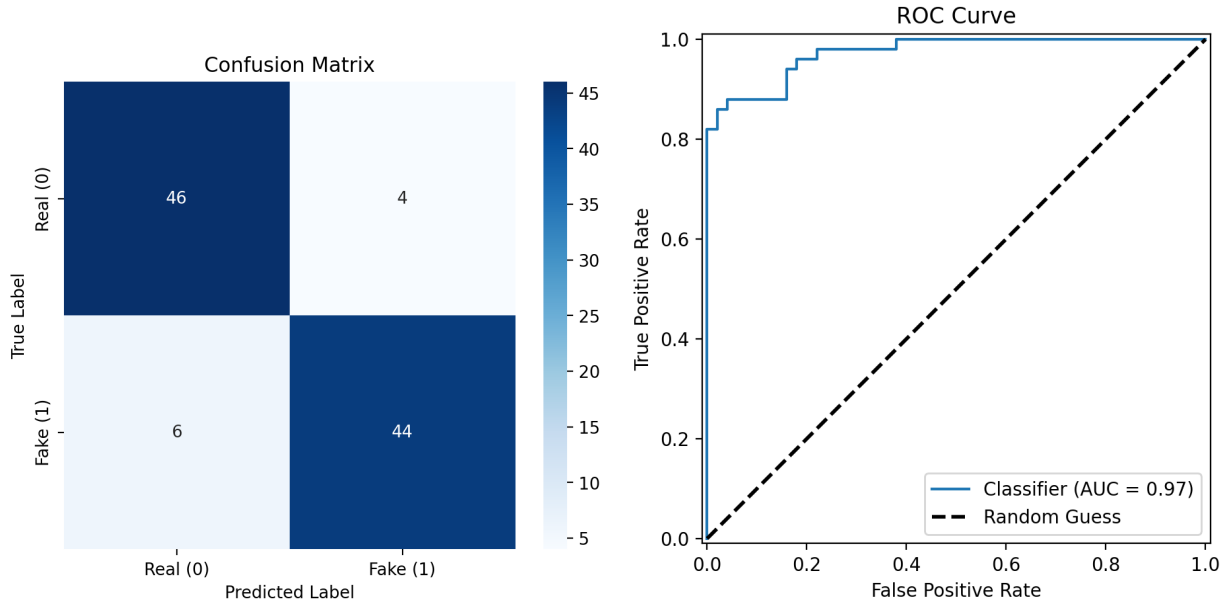


Figura 4.3: Matriz de confusión (izquierda) y curva ROC (derecha) para D^3 en el conjunto de validación personalizado.

entre ambas vistas. Este mecanismo tiende a enfatizar inconsistencias locales y señales forenses que se manifiestan de forma diferencial ante perturbaciones estructuradas. En imágenes sintéticas altamente coherentes a nivel local —característica cada vez más común en modelos generativos recientes— esas discrepancias pueden ser más sutiles, elevando falsos negativos. A la vez, el enfoque por discrepancia puede ser menos sensible a cues globales de "estilo" que afectan a fotografías reales antiguas (p. ej., grano, compresión), lo que contribuye a reducir falsos positivos frente a modelos que reaccionan ante dichas degradaciones.

En escenarios de verificación noticiosa, D^3 aparece como una alternativa robusta a UFD: ofrece un perfil ligeramente más conservador en falsas alarmas y mantiene capacidad discriminativa comparable. La decisión entre ambos modelos se justifica principalmente por el perfil FP/FN y por el umbral seleccionado, más que por restricciones de rendimiento. La tabla 4.2 documenta un par de clasificaciones erroneas (1 falso positivo y 1 falso negativo) cometidas por D^3, lo cual puede ser útil para análisis cualitativos posteriores sobre las características de las imágenes que generan confusión en el modelo.

4.4. Validación de resultados para los modelos de Hugging Face

Además de los modelos extraídos de literatura académica, se evaluaron tres detectores disponibles en Hugging Face que resultan atractivos por su accesibilidad y facilidad de integración (pesos y código listos para uso). En este proyecto, su rol es el de baselines prácticos: no se asumen equivalen-

Tipo	Imagen (ID/archivo)
Falso positivo	
Falso negativo	

Cuadro 4.2: Ejemplos de falso positivo (FP) y falso negativo (FN) cometidos por D^3 en el conjunto de validación personalizado.

tes en objetivo ni en cobertura a UFD o D^3, y por tanto sus métricas se interpretan principalmente como evidencia de transferencia al dominio noticioso.

Para realizar una comparación justa, los tres modelos recibieron imágenes con el mismo preprocesamiento que esperan sus procesadores originales: conversión a RGB y aplicación del *image processor* del modelo (`Transformers`), que incluye el re-escalado al tamaño de entrada y la normalización correspondiente. Adicionalmente, y como parte del aprendizaje por transferencia descrito en el Capítulo 3, se mantuvieron congelados los backbones y se entrenó únicamente la cabeza de clasificación. En inferencia, se utilizó el puntaje asociado a la clase positiva (`fake`) para construir las curvas ROC (usando las salidas directas del modelo, sin calibración posterior), y una decisión binaria a umbral fijo (por defecto, 0.5) para las matrices de confusión.

4.4.1. `Smogy/SMOGY-Ai-images-detector`

La Figura 4.4 presenta los resultados de `Smogy/SMOGY-Ai-images-detector`. En esta evaluación, el modelo alcanzó una exactitud de 0,7300, precisión de 0,7949, recall de 0,6200, F1 de 0,6966 y ROC-AUC de 0,8288.

El desglose de errores evidencia una tendencia a sub-detectar el contenido sintético: 19 falsos negativos frente a 8 falsos positivos. Este perfil es compatible con que `Smogy/SMOGY-Ai-images-detector` se entrene/afine como detector generalista sobre dominios heterogéneos (contenido de Internet, Kaggle y arte), donde la separación real/sintético puede depender en parte de estilos visuales y distribuciones de calidad distintas a las de fotografía de prensa histórica. Una interpretación plausible es que, al trasladarlo a un dominio noticioso con escenas variadas, compresión y estética editorial, su puntaje se active con mayor claridad en muestras sintéticas con huellas visuales más evidentes, mientras que pierda cobertura ante imágenes sintéticas altamente fotorrealistas o postprocesadas.

El ROC-AUC de 0,8288 indica que, aunque el umbral de decisión usado en la matriz de confusión

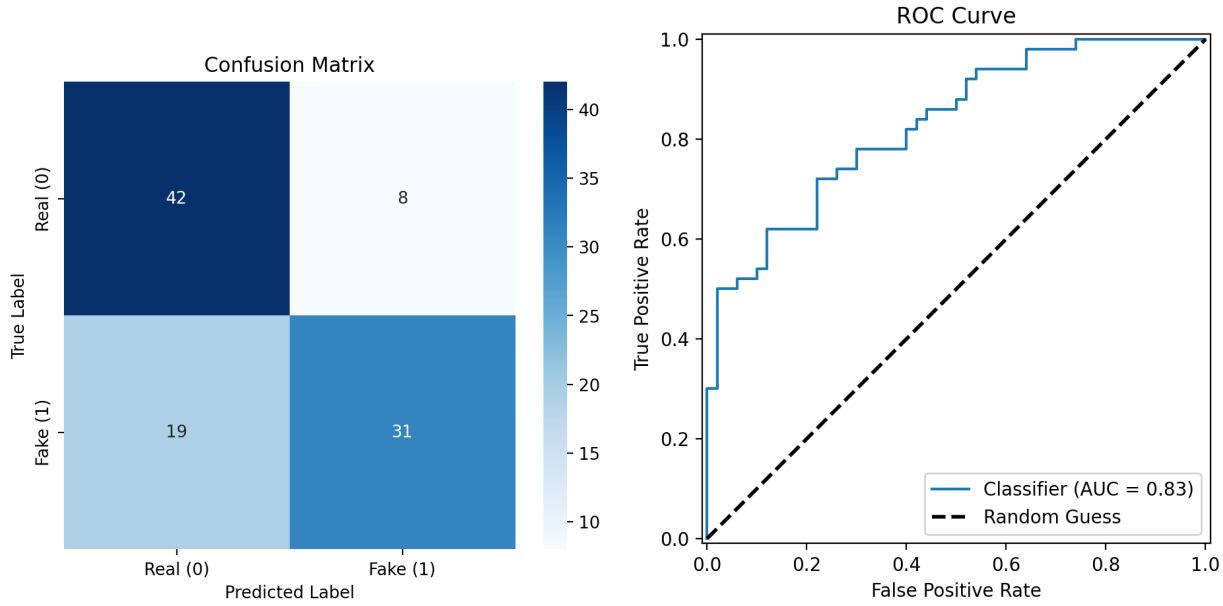


Figura 4.4: Matriz de confusión (izquierda) y curva ROC (derecha) para Smogy/SMOGY-Ai-images-detector en el conjunto de validación personalizado.

produce un recall moderado, la señal de puntaje conserva información discriminativa. Por tanto, una calibración de umbral sobre el dominio noticioso podría ajustar el compromiso precisión/recall según el objetivo operativo (reducir falsas alarmas vs aumentar cobertura). No obstante, la brecha respecto a UFD y D^3 sugiere que el límite principal no es únicamente de umbral, sino de transferencia: el modelo no se alinea tan bien con la distribución visual noticiosa y con generadores contemporáneos diversos.

4.4.2. `Ateeqq/ai-vs-human-image-detector`

La Figura 4.5 muestra el desempeño de `Ateeqq/ai-vs-human-image-detector`. El modelo obtuvo una exactitud de 0,6800, precisión de 0,9500, recall de 0,3800, F1 de 0,5429 y ROC-AUC de 0,7536.

El perfil de error es marcadamente conservador: sólo 1 falso positivo, pero 31 falsos negativos. Esta combinación explica su precisión extremadamente alta (cuando predice "fake", casi siempre acierta) y su recall bajo (omite la mayoría de imágenes sintéticas). Más que interpretarse como un "fracaso" general del detector, este patrón es compatible con su objetivo y datos de entrenamiento: `Ateeqq/ai-vs-human-image-detector` se orienta a distinguir imágenes humanas frente a imágenes generadas por modelos de última generación, con énfasis en ejemplos de alta calidad. Una hipótesis plausible es que, al trasladarlo a un dataset noticioso con escenas, objetos y contextos variados (no necesariamente centrados en rostros o en el tipo de "human vs IA" del que aprende), la frontera aprendida se vuelva más selectiva y se manifieste como infradetección sistemática.

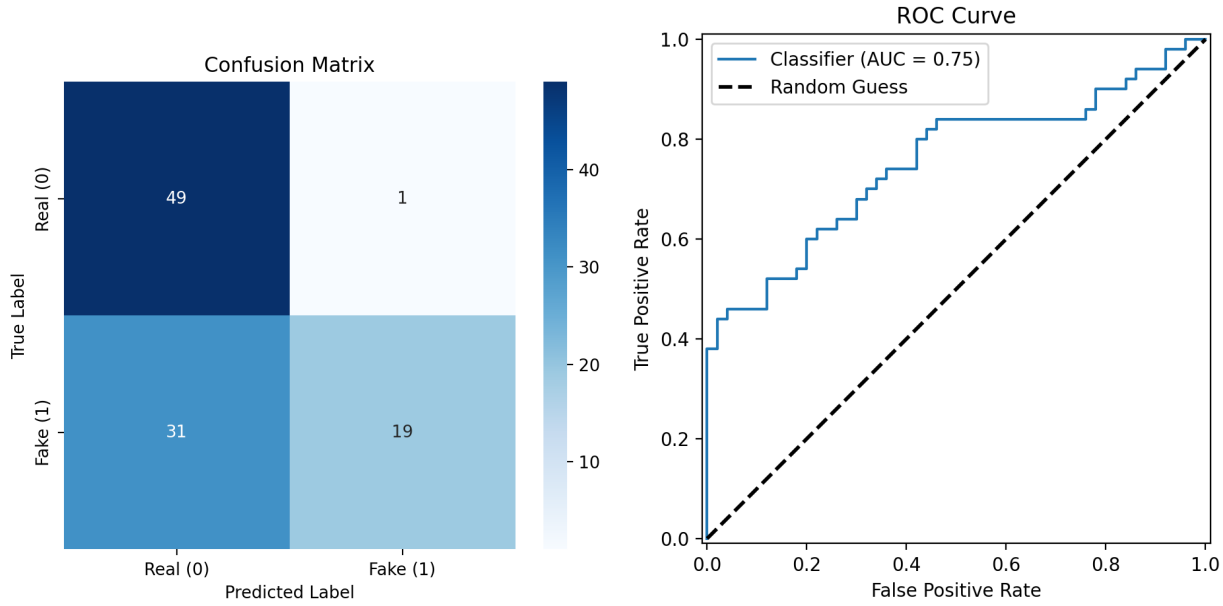


Figura 4.5: Matriz de confusión (izquierda) y curva ROC (derecha) para Ateeqq/ai-vs-human-image-detector en el conjunto de validación personalizado.

En términos de implicaciones, el modelo puede ser útil como verificador de alta precisión (p. ej., para priorizar revisiones humanas cuando marca "fake"), pero no como filtro exhaustivo. El ROC-AUC de 0,7536 sugiere que existe margen de mejora por ajuste de umbral, aunque el aumento de recall implicaría un incremento de falsos positivos. En este caso, el ajuste de umbral debe entenderse como una decisión operativa: el modelo parece priorizar minimizar acusaciones falsas, lo cual puede ser deseable en flujos periodísticos.

4.4.3. `prithivMLmods/deepfake-detector-model-v1`

Finalmente, la Figura 4.6 resume el desempeño de `prithivMLmods/deepfake-detector-model-v1`. El modelo alcanzó una exactitud de 0,7200, precisión de 1,0000, recall de 0,4400, F1 de 0,6111 y ROC-AUC de 0,8932.

La matriz de confusión revela un caso extremo de aversión a falsos positivos: el modelo no clasifica ninguna imagen real como sintética (0 falsos positivos), pero omite 28 de 50 imágenes sintéticas. Esta dinámica es compatible con su propósito y dominio de entrenamiento: `prithivMLmods/deepfake-detector-model-v1` es un detector especializado en deepfakes faciales, por lo que su señal tiende a ser más informativa cuando hay rostros dominantes y patrones de manipulación compatibles con los aprendidos. En un entorno noticioso, donde gran parte del contenido visual incluye escenas abiertas, objetos, eventos o planos no centrados en caras, esta especialización podría limitar la transferencia y reducir su cobertura como detector generalista.

Sin embargo, su ROC-AUC de 0,8932 es una señal importante: a pesar del umbral conservador

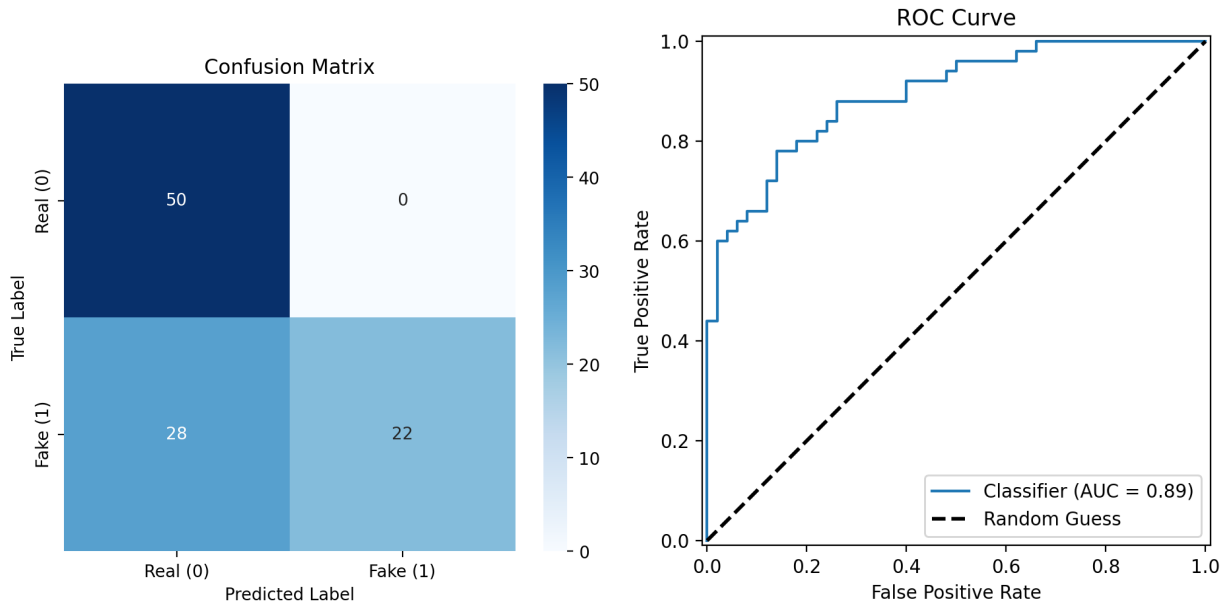


Figura 4.6: Matriz de confusión (izquierda) y curva ROC (derecha) para prithivMLmods/deepfake-detector-model-v1 en el conjunto de validación personalizado.

reflejado en la matriz de confusión, el modelo ordena razonablemente bien las imágenes reales y sintéticas. Esto sugiere que una recalibración explícita del umbral sobre el dominio noticioso podría recuperar parte del recall manteniendo una tasa baja de falsos positivos, haciendo viable su uso como componente especializado (por ejemplo, en piezas noticiosas donde la presencia de rostros sea dominante), más que como detector universal.

4.5. Comparación de resultados entre los modelos

La Tabla 4.3 resume el desempeño de todos los modelos evaluados sobre el conjunto de validación personalizado (100 imágenes, balanceadas). Dado este tamaño, cada error corresponde a un punto porcentual de exactitud; por ello, diferencias pequeñas deben interpretarse como evidencia inicial y complementarse con conteos de error y una noción explícita de incertidumbre estadística.

Para hacer explícita la evidencia empírica disponible con $n = 100$, la Tabla 4.4 reporta los conteos de la matriz de confusión (TP, FP, TN, FN) y un intervalo de confianza (IC) del 95 % para la *accuracy* (Wilson, aproximación binomial). Estos IC reflejan que, con este tamaño muestral, diferencias de 1–2 aciertos no sustentan un ranking definitivo: por ejemplo, los IC de UFD y D^3 se solapan ampliamente, por lo que el ordenamiento por *accuracy* debe entenderse como una tendencia observada, no como superioridad concluyente.

Complementariamente, la Tabla 4.5 reporta IC del 95 % para F1 y ROC-AUC. Para F1, el IC se estimó mediante bootstrap paramétrico sobre la matriz de confusión (re-muestreo multinomial, $B =$

Modelo	Accuracy	Precision	Recall	F1	ROC-AUC
UniversalFakeDetect	0,9100	0,9020	0,9200	0,9109	0,9740
D^3	0,9000	0,9167	0,8800	0,8980	0,9732
Smogy/SMOGY-Ai-images-detector	0,7300	0,7949	0,6200	0,6966	0,8288
prithivMLmods/deepfake-detector-model-v1	0,7200	1,0000	0,4400	0,6111	0,8932
Ateeqq/ai-vs-human-image-detector	0,6800	0,9500	0,3800	0,5429	0,7536
DistilDIRE	0,5400	0,5556	0,4000	0,4651	0,5680

Cuadro 4.3: Resumen de métricas en el conjunto de validación personalizado. La clase positiva corresponde a "fake" (imagen sintética).

Modelo	TP	FP	TN	FN	Errores	Accuracy (IC 95 %)
UniversalFakeDetect	46	5	45	4	9	0,910 [0,838; 0,952]
D^3	44	4	46	6	10	0,900 [0,826; 0,945]
Smogy/SMOGY-Ai-images-detector	31	8	42	19	27	0,730 [0,636; 0,807]
prithivMLmods/deepfake-detector-model-v1	22	0	50	28	28	0,720 [0,625; 0,799]
Ateeqq/ai-vs-human-image-detector	19	1	49	31	32	0,680 [0,583; 0,763]
DistilDIRE	20	16	34	30	46	0,540 [0,443; 0,634]

Cuadro 4.4: Conteos de aciertos y errores en el conjunto de validación personalizado (50 reales/50 sintéticas). La clase positiva es "fake". El IC del 95 % para *accuracy* se estima con el intervalo de Wilson.

200 000), lo cual proporciona una medida de variabilidad compatible con los conteos observados. Para ROC-AUC, en ausencia de los puntajes individuales por imagen en este capítulo, se reporta un IC asintótico aproximado basado en la varianza propuesta por Hanley–McNeil (asumiendo $n_+ = n_- = 50$). En ambos casos, los intervalos deben interpretarse como aproximaciones que enfatizan la incertidumbre asociada a $n = 100$.

En términos globales, UniversalFakeDetect y D^3 son los modelos con mejor desempeño observado en esta muestra: alcanzan *accuracy* cercana al 90 % y ROC-AUC alrededor de 0,97, lo que sugiere una separabilidad alta incluso frente a cambios de umbral. Sin embargo, dado el solapamiento de sus IC y el hecho de que una diferencia de 0,01 en *accuracy* equivale a un solo ejemplo, el resultado debe interpretarse como evidencia inicial de ventaja (consistente con su diseño), no como una conclusión categórica sobre un "mejor modelo".

La ventaja observada de estos detectores es compatible con su fundamento en representaciones de gran escala (CLIP) y con su estrategia de refinamiento con backbone congelado: en lugar de memorizar artefactos idiosincráticos de un generador, aprenden una frontera lineal (UFD) o una combinación de atención con señales de discrepancia (D^3) sobre un espacio semántico-estilístico generalista. Bajo esta lectura, es plausible que ello favorezca la transferencia a un conjunto de validación construido con generadores contemporáneos y variados.

La comparación también evidencia que el rendimiento no depende únicamente de "acertar más", sino del perfil de error (Tabla 4.4). UFD presenta un compromiso balanceado (FP=5, FN=4), mien-

Modelo	F1 (IC 95 %)	ROC-AUC (IC 95 %)
UniversalFakeDetect	0,9109 [0,8454; 0,9630]	0,9740 [0,9419; 1,0000]
D^3	0,8980 [0,8261; 0,9545]	0,9732 [0,9406; 1,0000]
Smogy/SMOGY-Ai-images-detector	0,6966 [0,5753; 0,7959]	0,8288 [0,7474; 0,9102]
prithivMLmods/deepfake-detector-model-v1	0,6111 [0,4667; 0,7333]	0,8932 [0,8282; 0,9582]
Ateeqq/ai-vs-human-image-detector	0,5429 [0,3896; 0,6753]	0,7536 [0,6584; 0,8488]
DistilDIRE	0,4651 [0,3256; 0,5895]	0,5680 [0,4556; 0,6804]

Cuadro 4.5: IC del 95 % para F1 y ROC-AUC en el conjunto de validación personalizado ($n = 100$). F1: bootstrap paramétrico sobre la matriz de confusión (multinomial, $B = 200\,000$). ROC-AUC: IC asintótico aproximado (Hanley–McNeil) asumiendo $n_+ = n_- = 50$.

tras que D^3 reduce levemente los falsos positivos a costa de aumentar los falsos negativos (FP=4, FN=6). En aplicaciones de verificación periodística, esta diferencia es relevante: los falsos positivos pueden traducirse en acusaciones indebidas sobre material auténtico (con costo reputacional y riesgo de censura), mientras que los falsos negativos implican permitir la circulación de imágenes sintéticas potencialmente desinformativas. Por tanto, la selección del modelo y el ajuste de su umbral deben alinearse con el riesgo operativo que se desea minimizar.

Los modelos de Hugging Face muestran un comportamiento heterogéneo y, en general, inferior a los detectores basados en CLIP. `Smogy/SMOGY-Ai-images-detector` ofrece el mejor equilibrio dentro de este grupo, pero aún sufre una tasa elevada de falsos negativos (FN=19), sugiriendo sensibilidad limitada a imágenes sintéticas hiperrealistas. En contraste, `Ateeqq/ai-vs-human-image-detector` y `prithivMLmods/deepfake-detector-model-v1` priorizan fuertemente evitar falsos positivos, logrando precisiones muy altas, pero sacrifican recall de manera pronunciada (FN=31 y FN=28, respectivamente). Este patrón revela un sesgo de cobertura: ambos modelos parecen detectar sólo un subconjunto de "fakes evidentes" (o compatibles con su dominio de entrenamiento), por lo que son más apropiados como detectores de alta precisión para priorización de casos que como filtros generales.

Finalmente, DistilDIRE se comporta de forma cercana al azar en este escenario. La distancia entre sus resultados y los reportes de la literatura en benchmarks de difusión sugiere una implicación relevante: los detectores que dependen de señales de reconstrucción/inversión y de supuestos sobre el mecanismo generativo pueden ser especialmente sensibles al cambio de dominio (nuevos generadores, compresión, degradaciones editoriales). En consecuencia, para el contexto de noticias reales vs imágenes sintéticas contemporáneas, los enfoques basados en representaciones generales y en señales universales (como UFD y D^3) muestran, en esta evaluación, un desempeño empírico más consistente.

4.5.1. Consideraciones éticas y de impacto social

Los resultados obtenidos no deben interpretarse como evidencia de que los modelos evaluados puedan sustituir el juicio humano en procesos de verificación periodística. Incluso los modelos con mejor desempeño presentan errores de clasificación que pueden generar consecuencias relevantes

dependiendo del contexto de uso.

Desde una perspectiva ética, los falsos positivos constituyen un riesgo importante, ya que una imagen auténtica podría ser catalogada incorrectamente como sintética. En escenarios periodísticos o de comunicación pública, este tipo de error podría afectar la credibilidad de periodistas, medios de comunicación, organizaciones o ciudadanos que compartan material legítimo. Asimismo, un uso automatizado de estas herramientas podría favorecer decisiones de moderación o censura basadas en clasificaciones incorrectas.

Por otra parte, los falsos negativos también representan una preocupación significativa. Una imagen sintética clasificada como auténtica puede contribuir a la difusión de desinformación visual, especialmente cuando el contenido se presenta como evidencia de eventos de interés público. Dado el creciente realismo alcanzado por los modelos generativos contemporáneos, este riesgo resulta particularmente relevante en contextos electorales, sociales o informativos.

En consecuencia, los detectores analizados deben entenderse como herramientas de apoyo para procesos de verificación y análisis forense, y no como mecanismos de decisión autónoma. Su utilización responsable requiere supervisión humana, validación adicional mediante otras fuentes de evidencia y una comunicación transparente de sus limitaciones y niveles de incertidumbre. Bajo esta perspectiva, el principal aporte de los modelos evaluados consiste en asistir la identificación temprana de contenido potencialmente sintético, reduciendo la carga operativa de los verificadores sin reemplazar la evaluación crítica realizada por expertos humanos.

4.6. Interfaz de usuario para la comparación de modelos

Como complemento a la validación cuantitativa presentada en este capítulo, se desarrolló una interfaz de usuario interactiva utilizando Streamlit para facilitar la comparación operativa de los modelos sobre el conjunto de validación personalizado. Es importante enfatizar que esta interfaz se concibe como un prototipo demostrativo (entregable complementario) y **no** como evidencia de adopción o utilidad en campo: el sustento del trabajo se basa en la comparación experimental y en las métricas reportadas, mientras que la interfaz aporta una forma práctica de inspeccionar salidas y comunicar resultados a usuarios (p. ej., jurados) de manera reproducible.

La implementación permite operar en dos modalidades: (i) evaluación individual de imágenes, donde el usuario puede cargar una imagen y obtener las predicciones de cada modelo junto con sus respectivas puntuaciones de confianza, y (ii) evaluación por lotes, donde se cargan imágenes organizadas en dos clases (“fake” y “real”) para clasificar un conjunto y visualizar un resumen comparativo. En ambos casos, la interfaz utiliza los checkpoints seleccionados posterior al refinamiento descrito en el Capítulo 3. Para la modalidad individual, se incluyó un modo de consenso por *votación mayoritaria* entre modelos; esta funcionalidad se incorpora con fines exploratorios y de usabilidad, y no fue validada como un ensamblado con garantías de mejora frente a los modelos por separado. La Figura 4.7 muestra una captura de pantalla de la interfaz desarrollada.

Debido a las complejidades técnicas relacionadas al uso del modelo DistilDIRE, su integración en la interfaz se realizó de forma condicional. En concreto, el modelo se ejecuta únicamente cuando el usuario provee un conjunto de imágenes que haya sido preprocesado previamente con

las transformaciones necesarias para el modelo. Esto se debe a que el preprocesamiento requerido (estimación de ε_0 mediante inversión DDIM) es computacionalmente intensivo y no es factible realizarlo en tiempo real dentro de la aplicación, especialmente para usuarios sin acceso a hardware especializado. Por ende, para utilizar este modelo los usuarios deben preprocesar con antelación las imágenes utilizando el script "compute_dire_eps.sh" para obtener los archivos de entrada requeridos por la arquitectura, guardar los archivos preprocesados en el directorio `datasets/` y por último cargar el conjunto de imágenes originales comprimido en la interfaz. En el Anexo 6.2, se incluyen los requisitos de ejecución, dependencias, estructura de carpetas esperada y limitaciones operativas de la interfaz.

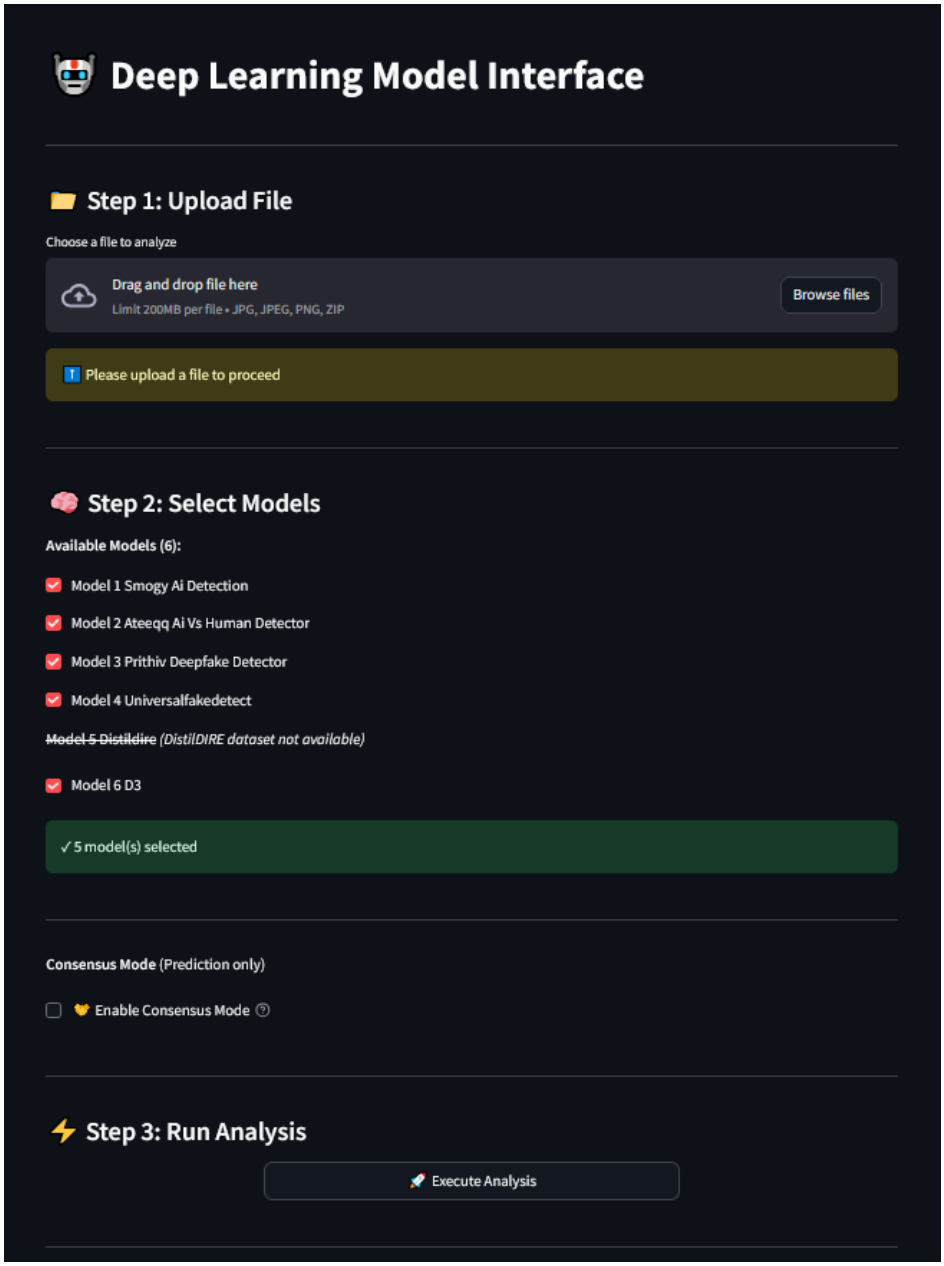


Figura 4.7: Interfaz de usuario para comparar modelos de detección de imágenes sintéticas.

Capítulo 5

Conclusiones

Este trabajo abordó el problema de la verificación de autenticidad visual en contexto noticioso, motivado por la creciente disponibilidad de imágenes generadas por IA y su uso potencial en redes sociales para apoyar narrativas de desinformación. En particular, se buscó contrastar el desempeño de distintos modelos de aprendizaje profundo para discriminar entre imágenes noticiables reales y sintéticas una vez adaptados mediante refinamiento.

En respuesta a la pregunta de investigación, los resultados de este proyecto sugieren que la comparación entre detectores se vuelve más consistente y útil (especialmente para toma de decisiones) cuando se cumplen tres condiciones: (i) los modelos se refinan bajo un protocolo homogéneo y controlado, (ii) la evaluación se realiza con métricas complementarias (matrices de confusión, precisión, recall, F1 y ROC-AUC) que permiten interpretar no sólo el acierto global sino también el tipo de error, y (iii) el conjunto de validación aproxima el escenario operativo objetivo, incluyendo variabilidad de generadores contemporáneos y características realistas del dominio periodístico. En este trabajo, dichas condiciones se abordaron al refinar los modelos seleccionados con MiRAGeNews bajo configuraciones comparables y evaluarlos en un conjunto de validación personalizado y balanceado (100 imágenes) diseñado para aproximar el ecosistema actual; por su tamaño, esta evaluación debe entenderse como evidencia externa inicial y controlada, no como una validación definitiva en campo.

La evidencia empírica obtenida (Tabla 4.3) permite aportar indicios, para este escenario y bajo este diseño experimental, sobre qué principios de modelado tienden a favorecer la transferencia al dominio noticioso. UniversalFakeDetect y D^3 alcanzaron los mejores desempeños (exactitudes de 0,9100 y 0,9000, respectivamente, y ROC-AUC cercanos a 0,97), mostrando una separabilidad robusta entre clases incluso ante variaciones de umbral. En este sentido, los resultados son consistentes con la idea de que los detectores basados en representaciones generales de gran escala (CLIP) y refinados con backbone congelado pueden transferir mejor cuando cambian los generadores y el estilo visual del dominio. En particular, entrenar únicamente cabezas de decisión ligeras puede reducir el riesgo de memorizar artefactos específicos del generador visto durante el refinamiento, lo cual es relevante en un escenario donde los modelos generativos evolucionan rápidamente.

Adicionalmente, el análisis por perfil de error refuerza que la selección del detector no debe basarse únicamente en la exactitud. UniversalFakeDetect presentó un equilibrio relativamente estable entre falsos positivos y falsos negativos, mientras que D^3 fue levemente más conservador al asignar la etiqueta “fake” (menos falsas alarmas a costa de omitir una fracción mayor de sintéticas). En verificación periodística, esta diferencia tiene implicaciones operativas: los falsos positivos pueden traducirse en cuestionamientos indebidos sobre material auténtico, mientras que los falsos negativos permiten que contenido sintético circule sin ser detectado. Por tanto, la decisión entre modelos (y la

calibración de su umbral) debe alinearse con el riesgo que se desea minimizar en el flujo de trabajo.

Los modelos disponibles en Hugging Face mostraron, en esta evaluación, que la transferencia al dominio noticioso depende del dominio de preentrenamiento y del perfil de decisión. `Smogy/SMOGY-Ai-images-detector` logró el mejor balance dentro de este grupo, pero con una infradetección notable de imágenes sintéticas de alta calidad (recall de 0,6200). En contraste, `Ateeqq/ai-vs-human-image-detector` y `prithivMLmods/deepfake-detector-model-v1` exhibieron perfiles altamente conservadores: precisiones muy altas (0,9500 y 1,0000) con recalls bajos (0,3800 y 0,4400). En consecuencia, estos modelos parecen más apropiados como verificadores de alta precisión para priorizar revisión humana cuando marcan "fake", que como filtros generales de cobertura amplia.

Finalmente, DistilDIRE obtuvo un desempeño cercano al azar (ROC-AUC de 0,5680), lo que sugiere un desacople entre la señal que explota el modelo y la distribución visual del conjunto de validación personalizado. Este hallazgo, junto con su preprocesamiento costoso (inversión DDIM para extraer ruido), indica que enfoques que dependen de supuestos específicos sobre el mecanismo generativo pueden ser particularmente vulnerables al cambio de dominio y menos viables operativamente en aplicaciones interactivas. En este sentido, los resultados del proyecto no sólo contrastan modelos por métricas, sino que también resaltan la necesidad de considerar restricciones de despliegue, latencia y complejidad de integración como criterios de selección.

Como cierre aplicado, la interfaz implementada en Streamlit demostró que es posible llevar esta comparación a un escenario de uso práctico, permitiendo evaluar imágenes individualmente o por lotes y visualizar el desempeño comparativo. En síntesis, bajo una evaluación externa controlada pero acotada, UniversalFakeDetect y D^3 mostraron mejor transferencia al dominio noticioso; sin embargo, estos hallazgos deben interpretarse como evidencia inicial y requieren ampliación del conjunto de evaluación, así como estudios adicionales de calibración (p. ej., umbrales) antes de considerarse una validación concluyente para su adopción operativa.

5.1. Limitaciones

- Las conclusiones de este trabajo deben interpretarse a la luz de varias limitaciones experimentales y de alcance. El conjunto de validación personalizado utilizado para la comparación contiene 100 imágenes balanceadas, lo cual permite un análisis controlado y comparable entre modelos, pero impone una incertidumbre apreciable: cada error altera la exactitud en un punto porcentual, por lo que diferencias pequeñas entre modelos pueden no ser significativas. Si bien se reportaron métricas complementarias e intervalos de confianza aproximados en la sección de resultados, en esta investigación no se realizaron múltiples corridas independientes de refinamiento por modelo (por ejemplo, con distintas semillas o particiones), ni un análisis formal de significancia que capture la variabilidad inducida por el entrenamiento; por tanto, la evidencia debe entenderse como comparativa bajo una única ejecución por configuración.

- Aunque el conjunto de validación fue diseñado para aproximar el ecosistema actual (con múltiples generadores contemporáneos), su construcción parte de un número acotado de contextos base y de una selección manual de imágenes reales. Esto puede introducir sesgos de selección

difíciles de cuantificar, así como dependencia de decisiones editoriales durante la curaduría. De manera particular, el pipeline de generación de imágenes sintéticas se apoya en prompts derivados (directa o indirectamente) de un LLM, y en algunos casos requirió iteraciones o ajustes para sortear restricciones de ciertos generadores. Estos dos factores pueden introducir sesgos sistemáticos (p. ej., homogeneidad semántica, estilo descriptivo característico del LLM o "prompt fingerprints"), de modo que parte de la separabilidad aprendida y evaluada podría reflejar regularidades del proceso de promptización más que artefactos de generación representativos de escenarios espontáneos o adversariales.

- Las imágenes reales utilizadas se tomaron de VisualNews, cuyas fuentes y periodo (principalmente 2010–2015) pueden diferir en compresión, resolución, estética y estilo fotográfico frente a material noticioso contemporáneo. Esta diferencia temporal puede inducir un cambio de distribución relevante: mientras las imágenes sintéticas evaluadas provienen de generadores modernos (con texturas y postprocesados actuales), las imágenes reales pueden exhibir patrones históricos de captura y publicación. En consecuencia, algunos falsos positivos/negativos podrían estar influenciados por esta asimetría ("real antiguo" vs "sintético moderno"), no únicamente por la autenticidad en sí.

- Aunque el problema y la motivación del trabajo se enmarcan en el contexto colombiano, la evidencia experimental no se construyó sobre un conjunto de datos específicamente colombiano (en términos de medios, temáticas locales, estilos de edición o distribución en redes sociales). Por ello, las conclusiones sobre aplicabilidad operativa en dicho contexto deben interpretarse con cautela, ya que la transferencia a subdominios geográficos puede verse afectada por factores de dominio no controlados.

- El refinamiento se realizó sobre MiRAGeNews utilizando únicamente la modalidad visual, pese a que el dataset es multimodal. Esto es coherente con el objetivo de evaluar detectores visuales, pero limita el alcance de las conclusiones: en entornos reales, señales textuales, metadatos, procedencia (provenance) y consistencia multimodal pueden aportar información adicional útil para la detección de desinformación.

- Finalmente, se priorizó un esquema de entrenamiento uniforme (mismo número de épocas y tamaños de batch) para preservar comparabilidad, lo cual reduce la exhaustividad en la búsqueda de hiperparámetros óptimos por modelo. En particular, es posible que algunos detectores requieran ajustes específicos de calibración (p. ej., umbral, técnicas de temperature scaling) o estrategias de adaptación de dominio para alcanzar su mejor desempeño en el contexto noticioso. Adicionalmente, la reproducibilidad completa de los resultados depende de la disponibilidad y preservación de artefactos experimentales (scripts de entrenamiento/evaluación, configuración exacta del entorno, semillas, listas de archivos y checkpoints refinados); si estos elementos no se publican o versionan de forma trazable, terceros podrían no replicar exactamente las cifras reportadas aun siguiendo la descripción metodológica.

5.2. Trabajo Futuro

En coherencia con las limitaciones descritas, las líneas de trabajo futuro se organizan por prioridad en tres frentes: (i) fortalecer la evidencia mediante una evaluación más amplia y auditada, (ii) calibrar y adaptar los modelos al punto operativo del dominio noticioso, y (iii) extender el sistema hacia señales multimodales y de procedencia, así como robustez avanzada.

1. El paso más urgente es robustecer la evaluación externa con conjuntos de validación y prueba más grandes y diversos, con imágenes reales más recientes, múltiples fuentes periodísticas y escenarios propios de redes sociales (reescalado, compresión agresiva, recortes, capturas de pantalla y re-posteo). Además de ampliar tamaño y diversidad, se propone incorporar una auditoría explícita del conjunto (criterios de curaduría, balance por tema/fuente, análisis de sesgos por antigüedad y compresión) para reducir asimetrías como "real antiguo" vs "sintético moderno". Dado el marco aplicado al contexto colombiano, es recomendable incluir material de medios colombianos y/o latinoamericanos (bajo licencias apropiadas), de modo que la evidencia de transferencia refleje subdominios geográficos relevantes. Idealmente, estas evaluaciones deben contemplar distribuciones de clase desbalanceadas que reflejen prevalencias reales y reportar métricas orientadas a costo (por ejemplo, curvas precisión–recall y análisis por puntos operativos), junto con pruebas de estrés ante postprocesados y generadores no vistos.

2. Una vez la evaluación sea más sólida, el siguiente paso es alinear la toma de decisiones con el riesgo operativo del flujo de verificación. Esto incluye: selección de umbrales por contexto (priorizar minimizar falsos positivos vs falsos negativos), calibración explícita de probabilidades (*temperature scaling* u otros métodos), y análisis de estabilidad del umbral bajo cambios de fuente, compresión o generador. En paralelo, es pertinente estudiar estrategias de adaptación de dominio con datos reales anotados localmente (noticias de medios colombianos), cuidando aspectos éticos y de licencia. Finalmente, se propone explorar ensamblajes o mecanismos de consenso que combinen detectores con perfiles complementarios (alta precisión vs alto recall) dentro de esquemas humano-en-el-bucle, donde las salidas del modelo se integren como triaje para priorizar revisión humana.

3. Con una base evaluativa y de calibración más madura, resulta natural extender el alcance más allá de la clasificación visual binaria incorporando señales multimodales (imagen + texto) y verificación de consistencia semántica entre imagen, titular y pie de foto. De manera complementaria, se propone integrar señales de procedencia (metadatos disponibles, trazabilidad de origen y, cuando aplique, estándares de autenticidad de contenido) para aproximar el escenario realista en el que la imagen rara vez aparece aislada. Por último, dada la fragilidad observada de DistilDIRE bajo cambio de dominio, es valioso investigar variantes que alineen mejor el componente de inversión/reconstrucción con el dominio objetivo (p. ej., difusores más cercanos a generadores contemporáneos o métodos de extracción de señal menos costosos), así como estudiar robustez ante postprocesados deliberados y ataques adaptativos.

En conjunto, estas líneas buscan transformar la evidencia inicial en una validación más concluyente, especificando primero la base evaluativa, luego la calibración del punto operativo y, finalmente, la ampliación hacia señales complementarias y despliegue robusto.

Capítulo 6

Anexos

6.1. Extracción de contexto de imágenes apoyado en LLMs

Para la construcción del conjunto de validación personalizado, particularmente para el split "fake", se utilizó un enfoque apoyado en LLMs para extraer información contextual (en forma de descripciones breves) a partir de 10 imágenes noticiables sintéticas seleccionadas aleatoriamente desde MiRAGeNews. Dado que este paso puede introducir sesgos (por el estilo descriptivo del modelo, por homogeneidad semántica o por inferencias no deseadas), se documenta de manera explícita y autocontenida para que se auditarlo sin depender de enlaces externos.

En concreto, este anexo incluye el prompt literal utilizado, el modelo (con fechas y parámetros de ejecución), los criterios de uso/edición aplicados cuando fue necesario ajustar el texto para cumplir términos y condiciones de generadores text-to-image, y una tabla con las salidas generadas y sus modificaciones (si aplicó).

6.1.1. Prompt, modelo y parámetros de ejecución

El objetivo del prompt fue obtener una descripción factual y concisa (máximo 30 palabras) que pudiera reutilizarse como prompt de generación en distintos modelos text-to-image. El prompt utilizado (transcrito literalmente) fue:

> Describe los contenidos de esta imagen en un parágrafo de 30 palabras o menos, detallando concisamente los hechos que se están retratando

El modelo LLM utilizado para esta tarea fue GPT-5.3-Instant, accedido a través de la interfaz web de ChatGPT (OpenAI) entre el 16 y el 18 de marzo de 2026. Se emplearon los parámetros predeterminados, con una temperatura de 1.0 y sin reintentos automáticos en caso de bloqueos por moderación o restricciones de contenido.

6.1.2. Criterios de uso y edición de prompts

Para cada una de las 10 imágenes, se registró la primera descripción devuelta por el LLM ("contexto inicial"). Cuando el contexto inicial incluía elementos que podían bloquear la generación por moderación o restricciones de ciertos modelos text-to-image (violencia o contenido explícito), se realizaron ajustes pequeños para eliminar referencias explícitas, manteniendo la esencia de la escena. El criterio operativo aplicado fue:

Para cada prompt que haya sido anulado debido a su descripción explícita por los modelos, se realizó la modificación de las palabras que se consideran más aludidas a contenido explícito o violento, y una vez se modificaron, se realizó un reintento de generación.

La Tabla 6.1 documenta, para cada imagen, el contexto inicial y la versión modificada (si aplica). "N/A" indica que no se realizaron modificaciones.

Cuadro 6.1: Contexto extraído de imágenes noticiables sintéticas utilizando LLMs.

Imagen	Contexto inicial	Modificación
	Un grupo de policías uniformados permanece alineado en un parque durante ceremonia solemne. Uno al frente inclina la cabeza mientras otros permanecen firmes, con público observando detrás de cuerda.	N/A
	Un hombre vestido de blanco saluda y sonríe desde un vehículo mientras una multitud lo rodea en una calle. Varias personas observan y banderas ondean durante lo que parece un acto público.	N/A

Imagen	Contexto inicial	Modificación
	Una mujer con bata blanca grita mientras varias personas la sujetan entre una multitud agitada. A su alrededor, gente observa y levanta teléfonos, registrando el momento en un entorno urbano concurrido.	Una mujer con bata blanca grita mientras varias personas la observan entre una multitud. A su alrededor, la gente observa y levanta teléfonos, registrando el momento en un entorno urbano concurrido.
	Un hombre con gorro y abrigo desgastado mira a la cámara mientras permanece frente a una gran multitud reunida en una plaza, donde se observan carpas, banderas y estructuras al fondo.	Un hombre con gorro y abrigo viejo mira a la cámara mientras permanece frente a una multitud reunida en una plaza, donde se observan carpas y estructuras urbanas al fondo.
	Un hombre sonriente toca una guitarra eléctrica frente a un micrófono durante una presentación al aire libre. Al fondo, otra persona y árboles sugieren un evento musical en vivo.	N/A

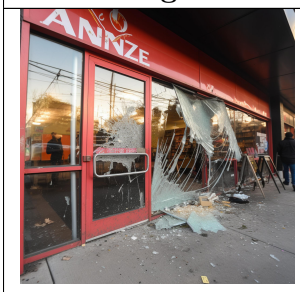
Imagen	Contexto inicial	Modificación
	Entrada de tienda con puertas y vitrinas rotas, vidrios fracturados y escombros que cubren la acera, mientras el interior queda expuesto, sugiriendo un acto de vandalismo o saqueo reciente.	Entrada de tienda con puertas y vitrinas fracturadas y escombros que cubren los alrededores, mientras el interior queda expuesto pero intacto.
	Un hombre trajeado habla con énfasis frente a varios micrófonos en una sala. Cámaras y periodistas lo rodean, sugiriendo una rueda de prensa o declaración oficial ante medios.	Un hombre con vestido formal habla con pasion frente a varios micrófonos en una sala. Hay cámaras y periodistas que lo rodean al ser una rueda de prensa de promoción de producto.
	Una explosión envuelve un vehículo en llamas mientras una multitud corre y se dispersa. Humo negro se eleva entre edificios, con escombros y caos dominando la escena urbana.	Un carro pasa mientras se toma una foto de un edificio evacuado. Hay humo que se eleva del edifico, y se percibe confusion de parte de los testigos del hecho urbano.

Imagen	Contexto inicial	Modificación
	Dos hombres con traje sonríen y estrechan manos frente a un edificio icónico junto al agua. La escena sugiere un encuentro formal o acuerdo en un entorno público reconocido.	N/A
	Una larga fila de personas espera en un andén mientras un tren permanece detenido. Los pasajeros, abrigados, aguardan abordar en una estación concurrida durante un momento de alta demanda.	N/A

Como respaldo adicional (no requerido para la auditabilidad de este anexo), se preservó un enlace a la conversación original utilizada durante la extracción, el cuál se puede acceder haciendo clic aquí.

6.2. Código fuente de la interfaz Streamlit para la comparación de modelos

Este anexo documenta los elementos necesarios para reproducir la interfaz desarrollada en Streamlit y evita interpretar su existencia como evidencia de adopción en campo: la interfaz se utiliza como prototipo demostrativo para inspeccionar predicciones y comparar de forma operativa los modelos (utilizando los checkpoints seleccionados tras el refinamiento).

- **Código fuente y punto de entrada:** El repositorio de la interfaz se puede encontrar haciendo clic aquí.
- **Dependencias:** El repositorio incluye un archivo `requirements.txt` con las dependencias necesarias para ejecutar la interfaz. Se recomienda crear un entorno virtual e instalar las dependencias utilizando el comando `pip install -r requirements.txt`.

- **Ejecución:** En la carpeta principal del proyecto, se puede iniciar la interfaz ejecutando el archivo `run_streamlit.bat` (Windows).
- **Estructura de carpetas y entradas:** La interfaz está diseñada para cargar imágenes (en formato PNG, JPG o JPEG) o archivos comprimidos (ZIP) que contengan imágenes. Los archivos comprimidos deben contener imágenes organizadas en subcarpetas que representen las clases (`real/`, `fake/`). La interfaz procesará los contenidos de los archivos de entrada y mostrará las predicciones de los modelos seleccionados.
- **Checkpoints utilizados:** La interfaz utiliza los checkpoints seleccionados posterior al refinamiento descrito en el Capítulo 3. Estos checkpoints deben estar ubicados en una carpeta respectiva dentro del directorio `models/`, donde dicha carpeta debe llevar el prefijo `mirage_`, seguido del número del modelo y su nombre (Por ejemplo: `mirage_model_1_resnet50`).
- **Modo consenso:** Se basa en votación mayoritaria entre modelos y se incluye con fines exploratorios; no se validó como un ensamblado optimizado ni se reportan métricas propias de este modo.

Bibliografía

[1] S. S. Baraheem and T. V. Nguyen, "AI vs. AI: Can AI Detect AI-Generated Images?," *Journal of Imaging*, vol. 9, p. 199, Sept. 2023.

[2] A. Lokner Lađević, T. Kramberger, R. Kramberger, and D. Vlahek, "Detection of AI-Generated Synthetic Images with a Lightweight CNN," *AI*, vol. 5, p. 1575–1593, Sept. 2024.

[3] A. Mahara and N. Rishe, "Methods and Trends in Detecting Generated Images: A Comprehensive Review," 2025.

[4] I. H. Sarker, "Ai-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems," *SN computer science*, vol. 3, no. 2, p. 158, 2022.

[5] L. González-Rodríguez and A. Plasencia-Salgueiro, "Uncertainty-Aware autonomous mobile robot navigation with deep reinforcement learning," in *Deep learning for unmanned systems*, pp. 225–257, Springer, 2021.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2021.

[7] U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models," 2024.

[8] Y. Lim, C. Lee, A. Kim, and O. Etzioni, "DistilDIRE: A Small, Fast, Cheap and Lightweight Diffusion Synthesized Deepfake Detection," 2024.

[9] Y. Yang, Z. Qian, Y. Zhu, O. Russakovsky, and Y. Wu, "D^3: Scaling Up Deepfake Detection by Learning from Discrepancy," 2025.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021.

[11] R. Huang, L. Dugan, Y. Yang, and C. Callison-Burch, "MiRAGeNews: Multimodal Realistic AI-Generated News Detection," Apr. 2025.

[12] A. Team, "Text-to-Image Arena — Photorealistic Cinematic Imagery," 2026.

[13] C. Drolsbach and N. Pröllochs, "Characterizing AI-Generated Misinformation on Social Media," *arXiv preprint arXiv:2505.10266*, 2025.

[14] C. Alvino, "Estadísticas de la situación digital de Colombia en el 2025," May 2025.

[15] OECD, *The OECD Truth Quest Survey: Methodology and findings*. No. 369, June 2024.

[16] P. Doke, P. Dhotre, and P. Kharat, "A Comprehensive Survey of Deepfake Detection Methods, Datasets, and Future Directions," in *2025 IEEE Pune Section International Conference (PuneCon)*, pp. 1–6, 2025.

[17] Tiwari, Manish R. and Patil, Sandip S., "Deepfake Image Detection: A Review of Existing Methods and a Hybrid CNN-Based Proposed Framework," *EPJ Web Conf.*, vol. 341, p. 01043, 2025.

[18] Z. Huang, T. Li, X. Li, H. Wen, Y. He, J. Zhang, H. Fei, X. Yang, X. Huang, B. Peng, and G. Cheng, "So-Fake: Benchmarking and Explaining Social Media Image Forgery Detection," 2025.

[19] Z. Huang, J. Hu, X. Li, Y. He, X. Zhao, B. Peng, B. Wu, X. Huang, and G. Cheng, "SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model," 2025.

[20] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, "Rethinking FID: Towards a Better Evaluation Metric for Image Generation," 2024.

[21] A. Visus, "Que es un Deep fakes, cómo se crean, cuáles fueron los primeros y su futuro," Jul 2021.

[22] F. J. García-Ull, "DeepFakes: The Next Challenge in Fake News Detection," *Anàlisi*, p. 103–120, Jun. 2021.

[23] E. Altuncu, V. N. L. Franqueira, and S. Li, "Deepfake: definitions, performance metrics and standards, datasets, and a meta-review," *Frontiers in Big Data*, vol. Volume 7 - 2024, 2024.

[24] J. Wu, W. Gan, Z. Chen, S. Wan, and H. Lin, "AI-Generated Content (AIGC): A Survey," 2023.

[25] R. D. Caballar and C. Stryker, "What is Computer Vision?," Nov 2025.

[26] A. Torralba, P. Isola, and W. Freeman, *Foundations of Computer Vision*. Adaptive Computation and Machine Learning series, MIT Press, 2024.

[27] W. E. Snyder and H. Qi, *Computer Vision, Some Definitions, and Some History*, p. 3–10. Cambridge University Press, 2017.

[28] D. L. T. Bale, L. C. Ochei, and C. Ugwu, "Deepfake Detection and Classification of Images from Video: A Review of Features, Techniques, and Challenges," *International Journal of Intelligent Information Systems*, vol. 13, no. 2, pp. 20–28, 2024.

[29] IBM, "What is image classification?," Sep 2025.

[30] S. Wang, "Research on Computer Vision Image Classification," *Academic Journal of Computing & Information Science*, vol. 6, no. 6, 2023.

[31] Y. Ma, B. Niu, and Y. Qi, "Survey of image classification algorithms based on deep learning," in *2nd International Conference on Computer Vision, Image, and Deep Learning* (B. H. bin Ahmad and F. Cen, eds.), vol. 11911, p. 119111X, International Society for Optics and Photonics, SPIE, 2021.

[32] Shaveta, "A review on machine learning," *International Journal of Science and Research Archive*, vol. 9, p. 281–285, May 2023.

[33] M. Chen, "What Is Supervised Learning?," Jul 2024.

[34] J. Murel and E. Kavlakoglu, "Aprendizaje por refuerzo," Mar 2024.

[35] ESIC, "¿Qué son y para qué sirven las redes neuronales artificiales? ," Aug 2025.

[36] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, p. 685–695, Sept. 2021.

[37] I. D. Mienye and T. G. Swart, "A Comprehensive Review of Deep Learning: Architectures, Recent Advances, and Applications," *Information*, vol. 15, no. 12, 2024.

[38] M. Iman, H. R. Arabnia, and K. Rasheed, "A Review of Deep Transfer Learning and Recent Advancements," *Technologies*, vol. 11, p. 40, Mar. 2023.

[39] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 820–836, 06 2024.

[40] IBM, "What is a loss function?," Jul 2024.

[41] Google, "Classification: ROC Curve and AUC," 2019.

[42] Smoggy, "AI Image Detector," 2024.

[43] A. Azam, "AI and Human Image Classification Model - v1," 2025.

[44] P. Sakthi, "Deepfake Detector Model - v1," 2025.

[45] U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models," 2025.

[46] Y. Yang, Z. Qian, Y. Zhu, O. Russakovsky, and Y. Wu, "D^3: Scaling Up Deepfake Detection by Learning from Discrepancy," 2025.

[47] A. Team, "Arena's Ranking Method," 2025.

[48] F. Liu, Y. Wang, T. Wang, and V. Ordonez, "Visualnews : A large multi-source news image dataset," 10 2020.