UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

TEMA:

Estudio metodológico de las herramientas tecnológicas actuales para detectar y reconocer DeepFakes, abordando la creciente amenaza de la manipulación de contenido multimedia generado por inteligencia artificial.

AUTOR:

Costa Mora, Brittany Valeria

Trabajo de Integración Curricular previo a la obtención del título de INGENIERO EN CIENCIAS DE LA COMPUTACIÓN

TUTOR:

Ing. Castro Aguilar, Gilberto Castro

Guayaquil – Ecuador

08 de septiembre del 2023

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

CERTIFICACIÓN

Certificamos que el presente trabajo de integración curricular fue realizado en su totalidad por la Srta. **Costa Mora, Brittany Valeria** como requerimiento para la obtención del título de **INGENIERO EN CIENCIAS DE LA COMPUTACIÓN.**

TUTOR (A)

f. _____

Ing. Castro Aguilar, Gilberto Castro

Guayaquil, a los ocho días del mes de septiembre del año 2023

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL
FACULTAD DE INGENIERÍA
CARRERA DE INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

DECLARACIÓN DE RESPONSABILIDAD

Yo, Costa Mora, Brittany Valeria

DECLARO QUE:

El Trabajo de Integración Curricular, **Estudio metodológico de las herramientas tecnológicas actuales para detectar y reconocer deepfakes, abordando la creciente amenaza de la manipulación de contenido multimedia generado por inteligencia artificial,** previo a la obtención del título de **INGENIERO EN CIENCIAS DE LA COMPUTACIÓN**.' ha sido desarrollado respetando derechos intelectuales de terceros conforme las citas que constan en el documento, cuyas fuentes se incorporan en las referencias o bibliografías. Consecuentemente este trabajo es de mi total autoría.

En virtud de esta declaración, me responsabilizo del contenido, veracidad y alcance del Trabajo de Integración Curricular referido.

Guayaquil, a los ocho días del mes de septiembre del año 2023

Brittany Costa

f. _____

Costa Mora, Brittany Valeria

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

AUTORIZACIÓN

Yo, **Costa Mora, Brittany Valeria**

Autorizo a la Universidad Católica de Santiago de Guayaquil a la **publicación** en la biblioteca de la institución del Trabajo de Integración Curricular, "**Estudio metodológico de las herramientas tecnológicas actuales para detectar y reconocer deepfakes, abordando la creciente amenaza de la manipulación de contenido multimedia generado por inteligencia artificial."**, cuyo contenido, ideas y criterios son de mi exclusiva responsabilidad y total autoría.

Guayaquil, a los ocho días del mes de septiembre del año 2023

EL AUTOR:

f. _____

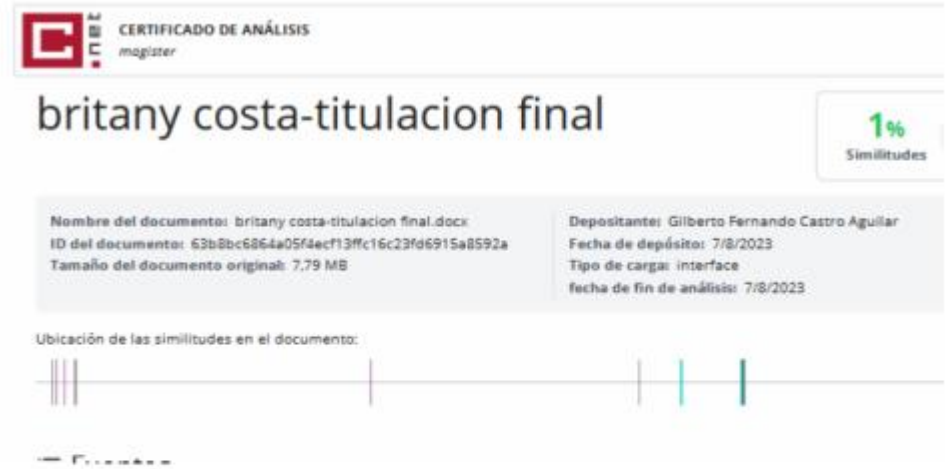
Costa Mora, Brittany Valeria

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

REPORTE ANTIPLAGIO



CERTIFICADO DE ANÁLISIS
magister

britany costa-titulacion final

1%
Similitudes

Nombre del documento: britany costa-titulacion final.docx
ID del documento: 63b8bc6864a05f4ecf13ffc16c23fd6915a8592a
Tamaño del documento original: 7,79 MB

Depositante: Gilberto Fernando Castro Aguilar
Fecha de depósito: 7/8/2023
Tipo de carga: interface
fecha de fin de análisis: 7/8/2023

Ubicación de las similitudes en el documento:

Fecha de elaboración: 30 de agosto 2023

Nombre del tutor: PhD. Gilberto Fernando Castro Aguilar, MSc.

Firma

Firma:

Ing. Gilberto Fernando Castro Aguilar, Mgs.
Tutor de Trabajo de Integración Curricular
Carrera de Ingeniería en Ciencias de la Computación

AGRADECIMIENTO

En la culminación de este arduo y gratificante trayecto académico, agradezco sinceramente a todos quienes contribuyeron de manera invaluable a este trabajo de titulación. Mi reconocimiento especial va hacia mi familia y amigos por su estímulo constante, comprensión y respaldo inquebrantable durante esta travesía académica. Sus palabras de ánimo fueron un faro en momentos desafiantes.

También extiendo mi gratitud a mi tutor, Gilberto Fernando Castro Aguilar, cuya experta orientación, paciencia y constante apoyo guiaron cada etapa de este proceso.

Finalmente, reconozco con gratitud a todas las fuentes de conocimiento y recursos que enriquecieron esta tesis. Cada autor, estudio y recurso utilizado ha aportado una dimensión única a mi comprensión y análisis.

Este logro no habría sido posible sin el apoyo, inspiración y dedicación de cada uno de ustedes. Su impacto perdurará en este trabajo de titulación y en mi desarrollo personal y académico.

Muchas gracias.

UNIVERSIDAD CATÓLICA
DE SANTIAGO DE GUAYAQUIL

FACULTAD DE INGENIERÍA

CARRERA DE INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

TRIBUNAL DE SUSTENTACIÓN

f. _____

ING. ANA CAMACHO CORONEL, MGS

DIRECTORA DE CARRERA

f. _____

ING. GALO CORNEJO GOMEZ, MGS

DOCENTE DE LA CARRERA

f. _____

ING. ROBERTO GARCÍA SANCHEZ, MGS

OPONENTE

ÍNDICE

ÍNDICE ..... **VIII**
RESUMEN ..... **IX**
ABSTRACT ..... **X**
INTRODUCCIÓN ..... **XI**
CAPÍTULO I **2**
Planteamiento Del Problema 2
Objetivos 5
Objetivo General 5
Objetivos Específicos 5
Alcances Del Problema 5
CAPITULO II **8**
Etiología 9
Tipos De DeepFake 11
Algoritmos Utilizados En La Creación De Deepfakes 18
Datasets Usados Por Herramientas De Reconocimiento Y Detección De Deepfakes 23
Herramientas De Reconocimiento Open Source 30
CAPÍTULO III **32**
Datasets 33
Métricas 34
Autores 35
CAPÍTULO IV **45**
Conclusiones **52**
Recomendaciones **54**

RESUMEN

Este trabajo de titulación busca abordar la creciente amenaza de los DeepFakes en la era digital, donde la manipulación de contenido multimedia generado por inteligencia artificial puede dificultar la distinción entre información real y falsa. El objetivo general de esta investigación ha sido identificar, analizar y evaluar las herramientas tecnológicas utilizadas para detectar y reconocer DeepFakes, con el propósito de enfrentar este desafío de manera efectiva.

Para lograrlo, se ha diseñado una ruta de trabajo que proporciona claridad en los procesos y pasos a seguir, permitiendo así la elaboración de una metodología sólida y bien fundamentada. La elección de la métrica AUC como criterio de evaluación ha permitido medir de manera global la capacidad discriminativa de las herramientas, asegurando una comparación objetiva y justa.

Los resultados obtenidos de la comparación y evaluación de cinco herramientas de detección de DeepFakes han revelado que Meso4 y Capsule se destacaron con los puntajes más altos en diferentes estudios, sobresaliendo especialmente en el conjunto de datos FaceForensics++. Además, se han identificado áreas de mejora y desafíos en la detección de DeepFakes, resaltando la importancia de considerar diferentes escenarios de generación de videos falsificados y la adaptabilidad de las herramientas a futuras técnicas de manipulación.

Este estudio aporta una valiosa contribución al campo de la detección de DeepFakes, ofreciendo una base sólida para futuras investigaciones en esta área. Los resultados y conclusiones obtenidos serán de gran interés para personas interesadas en la detección y mitigación de DeepFakes.

La implementación de estas herramientas de detección y la mejora continua de las mismas permitirá preservar la integridad y veracidad de los contenidos digitales en un contexto de rápida evolución tecnológica. De esta manera, se espera hacer frente a la amenaza de los DeepFakes y promover la confianza en la información en la sociedad actual.

Palabras Claves: *DeepFakes, inteligencia artificial, detección, AUC, aprendizaje automático, aprendizaje profundo.*

ABSTRACT

This thesis addresses the growing threat of DeepFakes in the digital age, where manipulating multimedia content generated by artificial intelligence can make distinguishing between accurate and false information difficult. The overall objective of this study was to identify, analyze, and evaluate the technological tools used to detect and recognize DeepFakes to address this challenge effectively.

A workflow has been designed to clarify the processes and steps to be followed, thus allowing the development of a solid and well-founded methodology. The choice of the AUC metric as the evaluation criterion has made it possible to globally measure the discriminative capacity of the tools globally, ensuring an objective and fair comparison.

The results obtained from comparing and evaluating five DeepFakes detection tools have revealed that Capsule stood out with the highest scores in different studies, excelling especially in the FaceForensics++ dataset. Furthermore, areas of improvement and challenges in DeepFakes detection have been identified, highlighting the importance of considering different scenarios of fake video generation and the adaptability of the tools to future manipulation techniques.

This research provides a valuable contribution to the field of DeepFake detection, offering a solid foundation for future research in this area. The results and conclusions obtained will significantly interest researchers, experts, and individuals interested in DeepFake detection and mitigation.

Implementing these detection tools and their continuous improvement will allow preserving the integrity and veracity of digital content in the context of rapid technological evolution. In this way, it is expected to address the threat of DeepFakes and promote trust in information today.

Key words: *DeepFakes, artificial intelligence, detection, AUC, machine learning, deep learning.*

INTRODUCCIÓN

Los medios digitales y los rápidos avances en inteligencia artificial han revolucionado la forma de producir, consumir y compartir información en nuestra sociedad actual. Sin embargo, en medio de esta revolución tecnológica, se cierne una amenaza oscura y formidable: los DeepFakes, los cuales pueden "fusionar, combinar, sustituir y superponer" diversas formas de medios de comunicación para producir fácilmente medios sintéticos que oculten la distinción entre información real y falsa (Kwok & Koh, 2021). El intercambio de rostros, especialmente en imágenes y vídeos, o la manipulación de la expresión facial se denominan métodos Deepfake (Mahmud & Sharmin, 2023). Su nombre se debe al uso de la tecnología de aprendizaje profundo, una rama del aprendizaje automático que aplica la simulación de redes neuronales a conjuntos de datos masivos. La inteligencia artificial aprende el aspecto de un rostro en diferentes ángulos para transponerlo a un objetivo, como si este llevara una máscara (Kerner & Risse, 2021).

Estas creaciones sintéticas han suscitado gran preocupación por la vulneración de la confianza, la distorsión de la realidad y la posibilidad de que se generalicen las campañas de desinformación. A medida que los DeepFakes siguen proliferando en diversas plataformas en línea, su creciente sofisticación plantea un reto importante a nuestra capacidad de discernir entre realidad y ficción, con implicaciones de gran alcance para las personas, las comunidades e incluso los procesos democráticos.

El presente trabajo de titulación busca profundizar en la investigación de la creciente amenaza de los DeepFakes, con el objetivo de esclarecer sobre sus alarmantes consecuencias, explorar las tecnologías subyacentes, y analizar las herramientas que existen para detectarlas y así, mitigar su impacto perjudicial.

Por ende, este trabajo de titulación está estructurado en cuatro capítulos cuyos contenidos se describen a continuación. En el capítulo I, se aborda el problema a resolver, se realiza un planteamiento general del problema, su ubicación en un contexto, se identifican causas y consecuencias, se delimita el problema, se formula y por último se realiza una evaluación de este. En el capítulo II, se detalla el marco teórico en donde se referencian definiciones conceptuales, normas, estándares, leyes, reglamentaciones, etc., las cuales se incluirán como anexo. En el tercer capítulo, se plantea la metodología de la investigación y los instrumentos a utilizar en este trabajo de titulación.

Por último, en el cuarto capítulo, se realizará un análisis de datos, las conclusiones que responden las interrogantes planteadas en los objetivos específicos, y las recomendaciones que deja el trabajo de titulación para que los lectores puedan emprender investigaciones similares o complementarias en el futuro.

CAPÍTULO I
EL PROBLEMA

Planteamiento Del Problema

Los algoritmos DeepFake han surgido como uno de los desarrollos más recientes y controversiales de la Inteligencia Artificial. Estos algoritmos utilizan el aprendizaje automático para generar contenidos realistas pero inventados, como imágenes, vídeos, audio y texto, a partir de un conjunto determinado de datos de entrada (Caporusso, 2021). La creciente prevalencia de los DeepFakes y su impacto en el panorama social y político suscitan gran preocupación, ya que tienen el potencial de desestabilizar gobiernos y manipular la opinión pública (Ahmed, 2021). Sin embargo, la influencia perjudicial de los DeepFakes va más allá de la política. Se infiltra en áreas de importancia social y económica, especialmente en el ámbito digital de las redes sociales donde se utilizan debido a su creciente accesibilidad de creación (Kietzmann et al., 2019) , donde las personas y las empresas son vulnerables a graves perjuicios.

Varios incidentes alarmantes relacionados con la tecnología DeepFakes sirven como recordatorios contundentes de su potencial destructivo en diferentes dominios. Por ejemplo, en el año 2019, una empresa energética británica fue víctima de una estafa de DeepFakes de voz, lo que provocó pérdidas financieras significativas. Los DeepFakes de voz aprovechan voces clonadas para hacerse pasar por personas de confianza por teléfono, explotando las relaciones personales y profesionales de las víctimas (Bateman, 2020). Otro caso notable fue el de un estudiante japonés que utilizó la tecnología DeepFakes para crear y distribuir vídeos pornográficos con la cara de una celebridad mediante un intenso proceso de entrenamiento. Este incidente supuso el primer caso penal en Japón relacionado con el uso malicioso de DeepFakes. Del mismo modo, en marzo de 2021, una madre de Pensilvania empleó la tecnología DeepFake para fabricar fotos y vídeos explícitos con el fin de atrapar a una animadora de instituto que era compañera de equipo de su hija (Gamage et al., 2021) . Estos incidentes evidencian el desconcertante potencial de los DeepFakes, que amplifican la

preocupación por la invasión de la privacidad, el daño a la reputación y la erosión de la confianza en diversos contextos sociales.

Los DeepFakes generados mediante algoritmos generativos avanzados de aprendizaje profundo han llegado a un punto en el que distinguir entre realidad e invención es cada vez más difícil. Desde su descubrimiento, se ha hecho evidente la facilidad con la que la tecnología DeepFakes puede emplearse con fines poco éticos y maliciosos, como difundir información errónea, suplantar la identidad de líderes políticos y difamar a personas inocentes. Desde entonces, los DeepFakes han experimentado avances significativos (Mirsky & Lee, 2021).

Sin embargo, en medio de las controversias que rodean a los DeepFakes, también han demostrado potencial para aplicaciones positivas. La comunidad científica ha empezado a explorar los beneficios de esta tecnología, especialmente en el ámbito educativo, donde los DeepFakes pueden mejorar las experiencias de aprendizaje mediante la incorporación de personajes familiares, creando así un entorno personalizado y atractivo. Dado que la tecnología DeepFake se encuentra aún en sus primeras fases, es crucial explorar ideas novedosas que pongan de relieve sus aspectos positivos, al tiempo que prosigue el debate en curso sobre sus futuras aplicaciones.

Un ámbito específico en el que los DeepFakes pueden resultar inmensamente valiosos es el de los entornos familiares remotos. Aprovechando esta tecnología, se pueden crear libros de cuentos digitales para niños, permitiendo a los abuelos asumir el papel de lectores incluso a distancia, reforzando así los lazos familiares a pesar de la separación física. Además, los DeepFakes pueden generar contenido para obituarios, sirviendo para celebrar la vida y los recuerdos de las personas fallecidas, proporcionando consuelo y apoyo a los seres queridos durante el proceso de duelo. Por otra parte, los vídeos DeepFake tienen el potencial de permitir la interacción con figuras contemporáneas o históricas destacadas, como científicos, políticos y artistas. Estas réplicas digitales ayudan a preservar y perpetuar su legado, permitiendo a los individuos interactuar con su trabajo y sus ideas. A través de los

DeepFakes, la influencia y las contribuciones de estas figuras pueden hacerse accesibles a las generaciones futuras (Caporusso, 2021).

Aunque la exploración de las posibilidades positivas de esta tecnología puede allanar el camino para su uso responsable y beneficioso en diversos entornos, persisten las consideraciones y preocupaciones éticas en torno a los DeepFakes.

A pesar de su naturaleza aparentemente inofensiva como herramienta de entretenimiento, la creación de DeepFakes puede tener consecuencias más amplias y preocupantes. En la esfera política, los DeepFakes pueden difundir información falsa y dañar la reputación de personas o partidos políticos enteros.

A mediados del 2020, varios DeepFakes se hicieron virales en Estados Unidos, muchos de ellos relacionados con la política. Ejemplos notables son los vídeos en los que el presidente Obama insulta al presidente Trump y Mark Zuckerberg admite que Facebook manipula a sus usuarios. Además, el Comité Nacional Demócrata (DNC) utilizó un DeepFake en el que aparecía su presidente, Tom Pérez, disculpándose por no asistir a una importante convención. El vídeo concluía con un descargo de responsabilidad que indicaba su condición de DeepFake, lo que demuestra las amenazas potenciales que plantean los DeepFakes en las elecciones presidenciales de 2020 (Ahmed, 2021).

Actualmente, una parte significativa de los DeepFakes que se encuentran en Internet son pornográficos, y las personas implicadas rara vez dan su consentimiento para su creación y difusión. Los DeepFakes exponen a cualquier persona con presencia en línea a la victimización (Karasavva & Noorbhai, 2021). Los DeepFakes de naturaleza sexual representan sólo un aspecto de la cosificación de la mujer en la cultura digital y visual. Estas manipulaciones no sólo despojan de agencia y autonomía a las mujeres implicadas, tanto a la persona cuyo rostro se inserta en el vídeo como a la actriz adulta cuyo cuerpo se altera y hace circular, sino que también contribuyen a un patrón más amplio de cosificación y explotación (Gosse & Burkell, 2020).

Además, los DeepFakes pueden afectar significativamente a la reputación de personas y empresas dentro de las redes sociales. Un caso ilustrativo es el de la periodista

de investigación india Rana Ayyub, que fue víctima de un DeepFake pornográfico en abril de 2018. En el vídeo, su rostro era sustituido por el de una actriz más joven con el cabello diferente. El vídeo se hizo rápidamente viral en toda la India, sembrando dudas sobre la credibilidad de Ayyub como periodista (Kerner & Risse, 2021). Actualmente se emplean diversas herramientas y algoritmos, como sistemas de autenticidad de imágenes y vídeos, herramientas de verificación de fuentes y herramientas de análisis de voz, para el reconocimiento y la detección de DeepFakes, con el objetivo de analizar la coherencia del contenido, autenticar datos, rastrear orígenes y detectar alteraciones de sonido.

El desarrollo constante de nuevas herramientas y algoritmos de detección sigue siendo crucial en la batalla contra los DeepFakes.

Objetivos

Objetivo General

El objetivo de este trabajo de titulación académica es identificar, analizar y evaluar las actuales herramientas tecnológicas utilizadas para detectar y reconocer DeepFakes, con el fin de abordar la creciente amenaza de la manipulación de contenido multimedia generado por inteligencia artificial.

Objetivos Específicos

- Identificar y caracterizar el funcionamiento de los métodos utilizados para crear deepfakes, aplicando técnicas de entrenamiento de algoritmos de aprendizaje profundo.
- Comparar las técnicas y algoritmos utilizados para la detección de deepfakes, identificando artefactos de compresión, análisis de movimiento, análisis de audio y detección de anomalías en la distribución de pixeles.
- Demostrar la funcionalidad de la herramienta tecnológica identificada y analizada sobre contenido multimedia generado por inteligencia artificial.

Alcances Del Problema

- El proyecto se llevará a cabo en un plazo de 16 semanas.

- Analizar y evaluar las tecnologías y métodos disponibles para identificar y prevenir la manipulación de contenido multimedia generado por inteligencia artificial.
- Revisión de lo que ya se ha escrito sobre el tema y la evaluación práctica de las herramientas y técnicas que se usan para detectar y reconocer deepfakes.
- Proporcionar una comprensión más profunda de las herramientas y técnicas disponibles para detectar y reconocer deepfakes, lo que puede ayudar a mejorar la seguridad y la confianza en el contenido multimedia en la era de la inteligencia artificial.

Justificacion E Importancia

La motivación para llevar a cabo este trabajo de titulación sobre la creciente amenaza de los deepfakes surge de la necesidad de abordar las importantes implicaciones sociales y éticas de esta tecnología emergente. Los DeepFakes tienen el potencial de erosionar la confianza, manipular la información y perturbar ámbitos críticos como la política, las redes sociales y la economía. Es imperativo comprender la naturaleza de los DeepFakes, su impacto potencial y desarrollar estrategias eficaces para detectar y mitigar sus efectos nocivos. De este modo, podremos salvaguardar la integridad de la información y proteger a las personas, comunidades e instituciones de las consecuencias perjudiciales de la manipulación de DeepFakes.

El objetivo de esta investigación es contribuir al conocimiento existente sobre DeepFakes, así como concienciar sobre los riesgos y retos que conllevan. Mediante una exploración profunda de sus fundamentos tecnológicos y un análisis exhaustivo de su impacto social, este estudio expondrá soluciones viables para la detección y prevención de los DeepFakes. Las ideas y conclusiones derivadas de esta investigación servirán para informar a responsables políticos, organizaciones y particulares, lo que permitirá comprender mejor las crecientes amenazas que plantean los DeepFakes y facilitará el desarrollo de contramedidas eficaces.

Además, dado que la tecnología DeepFake sigue evolucionando a un ritmo vertiginoso, es esencial explorar de forma proactiva las posibilidades positivas que puede ofrecer. Al esclarecer las posibles aplicaciones beneficiosas, como en la educación o la conservación de patrimonio, este trabajo de titulación pretende contribuir al discurso actual en torno al uso responsable y ético de la tecnología DeepFake.

En resumen, este trabajo busca ofrecer un análisis exhaustivo de la creciente amenaza que suponen los DeepFakes, haciendo hincapié en sus implicaciones sociales y en la urgente necesidad de adoptar medidas eficaces para hacer frente a este fenómeno. Al abordar esta cuestión oportuna y crítica, pretendemos contribuir a una comprensión más amplia de los DeepFakes y trabajar para establecer un paisaje digital más resistente y digno de confianza.

CAPITULO II

MARCO TEÓRICO Y CONCEPTUAL

Este marco teórico aborda el estudio y análisis de las técnicas utilizadas en la generación de deepfakes, así como los conceptos de identificación y caracterización definidos por (ASALE & RAE, n.d.-b) como la aportación de los datos personales necesarios para el reconocimiento y la determinación de atributos distintivos para diferenciar a alguien o algo de otros (ASALE & RAE, n.d.-a).

El notable fenómeno de los DeepFakes ha surgido debido a los rápidos avances tecnológicos, que conducen a una inteligencia artificial (IA) cada vez más sofisticada. Los DeepFakes, que combinan "deep learning" (aprendizaje profundo) y "fake" (falso), han cautivado el interés de los académicos y de la sociedad. Estas complejas creaciones mediáticas sintéticas, generadas mediante algoritmos de IA, pueden engañar a los observadores presentando contenidos auténticos.

La tecnología DeepFake, que incluye la manipulación de imágenes y vídeos, tiene sus raíces en el campo de la IA, que trata de comprender los procesos de pensamiento y el comportamiento humanos. El aprendizaje automático, un subconjunto de la IA, se emplea específicamente para permitir que los sistemas aprendan de los datos disponibles.

Los DeepFakes se basan en la convergencia de redes neuronales artificiales avanzadas, una rama crucial del aprendizaje automático, y el intrincado funcionamiento de la percepción humana. Al centrarse en la manipulación y síntesis de imágenes, los DeepFakes desafían nuestra comprensión fundamental de la autenticidad, la verdad y los límites de la percepción visual (Mirsky y Lee, 2021).

El aprendizaje profundo, una potente técnica de modelado que utiliza la composición de múltiples funciones no lineales, es una herramienta valiosa para establecer la estrecha relación entre las características de entrada y las etiquetas (Fan et al., 2021). El aprendizaje profundo se enmarca en el campo más amplio de la informática conocido como machine learning (aprendizaje automático), que tiene el potencial de revolucionar varias disciplinas, incluidas las ciencias epidemiológicas (Bi et al., 2019). El aprendizaje automático, comúnmente llamado ML, abarca el estudio científico de algoritmos y modelos estadísticos que permiten a los sistemas informáticos realizar tareas sin programación explícita. Una ventaja notable del ML es su capacidad para ejecutar de forma autónoma tareas designadas una vez que los algoritmos han aprendido de los datos disponibles.

En el contexto del procesamiento de imágenes, el ML desempeña un papel crucial al permitir la extracción de información significativa a partir de datos visuales. Los algoritmos de ML han avanzado significativamente las técnicas de procesamiento de imágenes, incluido el reconocimiento de objetos, la segmentación de imágenes y el análisis de patrones (Mahesh, 2019). El aprendizaje profundo, en particular, es famoso por su capacidad para representar datos complejos y de alta dimensión. Los autocodificadores profundos, una variante de las redes profundas, exhiben esta capacidad y se han empleado ampliamente en tareas como la reducción de la dimensionalidad y la compresión de imágenes.

La aparición de los DeepFakes como fenómeno intrigante en los últimos años ha aprovechado las capacidades de los algoritmos de aprendizaje profundo y aprendizaje automático.

Etiología

El fenómeno de las DeepFakes se originó con el desarrollo del software FakeApp por parte de un usuario anónimo de Reddit a finales de 2017, desatando la polémica al compartir vídeos pornográficos en los que supuestamente aparecían conocidas celebridades como Taylor Swift, Scarlett Johansson, Aubrey Plaza, Gal Gadot y Maisie Williams. Estos vídeos engañosos se crearon utilizando una técnica de vanguardia basada en el aprendizaje profundo, captando una importante atención mediática y propagándose rápidamente por

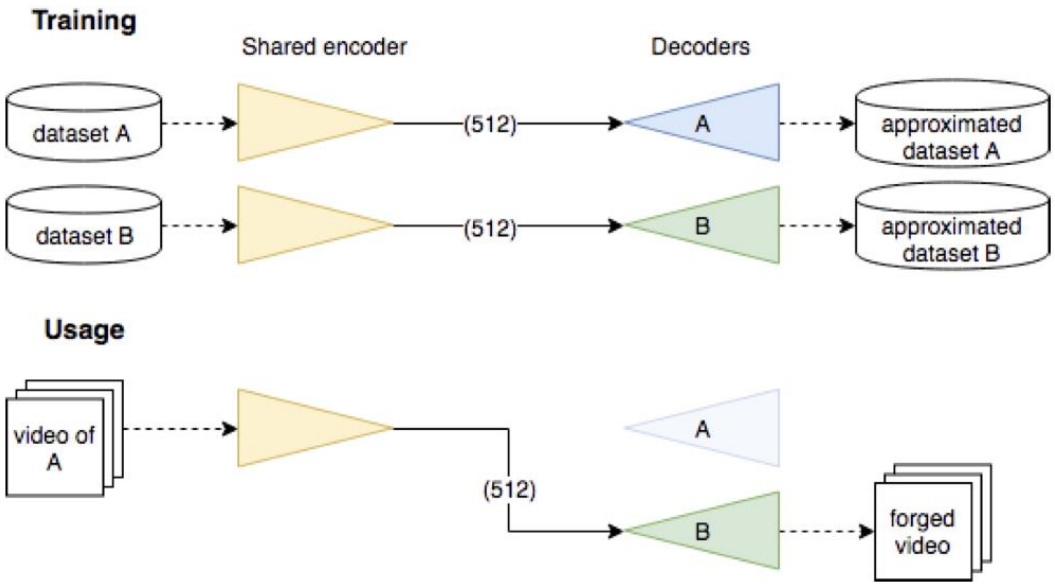
diversos foros en línea. Sin embargo, debido a la naturaleza controvertida del contenido, estos foros fueron finalmente prohibidos el 7 de febrero de 2018. Importantes plataformas como Discord, Gfycat y Twitter tomaron medidas decisivas para erradicar la presencia de DeepFakes. Desafortunadamente, a pesar de estos esfuerzos, la proliferación de DeepFakes continuó sin cesar, convirtiéndose en un problema global generalizado.

El origen de la técnica DeepFake se remonta a un ingeniero de software que publicó un kit de desarrollo, aprovechando la potencia de las herramientas de código abierto proporcionadas por líderes del sector como NVidia y Google. Este lanzamiento fundamental democratizó el acceso a la técnica, aunque requería un cierto grado de competencia técnica. La verdadera gravedad de la amenaza que representan los DeepFakes se hizo evidente cuando la Agencia de Proyectos de Investigación Avanzada de Defensa (DARPA) reconoció que incluso las personas sin conocimientos técnicos avanzados podían manipular fácilmente los medios visuales. En un incidente concreto de julio de 2017, investigadores de la Universidad de Washington crearon un vídeo falso del expresidente Obama, que sirvió como llamada de atención para concienciar sobre los peligros inherentes asociados a los DeepFakes. Posteriormente, en mayo de 2018, surgió un vídeo DeepFake de baja calidad en el que aparecía el presidente Donald Trump instando a los belgas a retirarse del Acuerdo de París sobre el cambio climático. Este penoso incidente puso aún más de relieve la naturaleza en constante evolución de la tecnología y su formidable capacidad para engañar y confundir al público desprevenido (Albahar & Almalki, 2019).

FakeApp, que utiliza el marco de correspondencia codificador-decodificador, es una herramienta única para crear DeepFakes. El proceso implica alimentar grandes cantidades de datos al sistema para entrenar el modelo para generar vídeos falsos convincentes. La cara de la persona de origen se inserta perfectamente en el vídeo de destino, creando una representación realista pero fabricada (Mahmud & Sharmin, 2023). En este método, el autocodificador extrae características latentes de las imágenes faciales, y el decodificador se emplea para reconstruir las imágenes faciales. Se requieren dos pares de codificador-decodificador para intercambiar rostros entre las imágenes de origen y de destino, cada uno

de los cuales se entrena con un conjunto de imágenes, y los parámetros del codificador se comparten entre los dos pares de redes (T. Nguyen et al., 2019).

Ilustración 1



Nota. La sección superior que muestra los componentes de entrenamiento con un codificador compartido resaltado en amarillo. En cambio, la sección inferior muestra la parte de utilización, en la que las imágenes de A se descodifican utilizando el descodificador de B (Afchar et al., 2018).

Los rápidos avances en el aprendizaje profundo y el aprendizaje automático, junto con la accesibilidad de herramientas como FakeApp, exigen un examen crítico de las implicaciones éticas y sociales asociadas con las falsificaciones profundas. A medida que estas tecnologías continúan avanzando, navegar por el complejo panorama de la autenticidad, la verdad y la percepción visual se vuelve esencial, asegurando un uso responsable y mitigando el daño potencial.

Tipos De DeepFake

Es importante resaltar que existen diferentes tipos de deepfakes de vídeo, cada uno con características y objetivos distintos. Estos tipos abarcan el intercambio de rostros, la

sincronización labial, el maestro de marionetas, la síntesis facial y la manipulación de atributos, así como los deepfakes de audio.

Intercambio De Rostros (Faceswap). El objetivo principal de las técnicas de intercambio facial (face-swapping) es sustituir a la perfección los rasgos faciales de un individuo objetivo por los de un individuo fuente, creando imágenes sintéticas que no se distinguen de las auténticas (Huang et al., 2023). Estos sofisticados DeepFakes se emplean con frecuencia para socavar la popularidad o la reputación de figuras prominentes mediante la fabricación de escenarios en los que nunca participaron realmente. Un ejemplo de este uso malicioso es la difusión de pornografía no consentida, que puede causar un daño significativo a la posición pública y el bienestar de una persona (Masood et al., 2021).

Los algoritmos de intercambio facial están diseñados para generar imágenes manipuladas que poseen la identidad visual y los atributos faciales de una persona diferente, abarcando factores como la pose, la expresión, la iluminación y el fondo. Para lograrlo, los desarrolladores suelen emplear dos enfoques principales: técnicas basadas en la fuente y técnicas basadas en el objetivo.

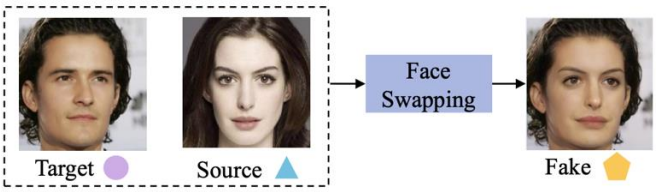
En el enfoque basado en el objetivo, el algoritmo extrae la "identidad" de la persona de la imagen de origen. La integra en la imagen de destino, conservando las características presentes en esta última. En cambio, el enfoque basado en la fuente consiste en modificar la imagen de origen en función de los atributos extraídos de la imagen de destino. Sin embargo, el inconveniente del enfoque basado en la fuente reside en la falta de control sobre los factores ambientales presentes en la imagen resultante (Walczyna & Piotrowski, 2023).

Al adoptar el enfoque basado en el objetivo, los algoritmos de face-swapping pueden transferir eficazmente la identidad del individuo de origen al individuo de destino, lo que resulta en una fusión perfecta y realista de los rasgos faciales. Este método permite una mayor precisión y control sobre el resultado final, ya que el algoritmo conserva los atributos inherentes de la imagen de destino a la vez que incorpora la identidad facial deseada de la imagen de origen. Por el contrario, el enfoque basado en la fuente puede dar lugar a

incoherencias y distorsiones, ya que el algoritmo modifica la imagen de origen sin tener en cuenta la información contextual de la imagen de destino.

Ilustración 2

Face Swapping



Nota. La cara de destino se sustituye por la cara de origen mediante el intercambio de caras para generar una cara falsa. En apariencia, la cara falsa se parece a la cara de origen en lugar de a la cara de destino (Huang et al., 2023).

Sin embargo, es importante señalar que ambos enfoques tienen ventajas y limitaciones. Aunque el enfoque basado en el objetivo ofrece un mayor control sobre el resultado final, depende en gran medida de la disponibilidad de imágenes objetivo de alta calidad que representen con precisión las características deseadas. Además, el enfoque basado en el objetivo puede tener dificultades para generar una imagen facial convincente en situaciones en las que la imagen de origen contiene información facial limitada o incompleta.

Por otro lado, el enfoque basado en el origen ofrece más flexibilidad en cuanto a las imágenes de origen disponibles, ya que puede utilizar una amplia gama de atributos faciales presentes en la imagen de origen para modificar el rostro de destino. Sin embargo, este enfoque puede provocar alteraciones involuntarias en el fondo de la imagen o en otros elementos no faciales, lo que puede dar lugar a intercambios de rostros menos realistas y visualmente incoherentes.

Por lo tanto, es fundamental que los desarrolladores conozcan en profundidad los puntos fuertes y débiles de cada enfoque a la hora de implementar algoritmos de intercambio de caras. Este conocimiento les permite tomar decisiones informadas y encontrar un equilibrio

entre la generación de resultados visualmente convincentes y realistas y la minimización de distorsiones e incoherencias no deseadas en el resultado final (Walczyna & Piotrowski, 2023).

En el campo de los DeepFakes, han surgido varios algoritmos y soluciones de software para facilitar el intercambio de caras basado en métodos de aprendizaje profundo. Algunos ejemplos notables son Faceswap y DeepFaceLab. Faceswap, presentado en 2017, emplea autocodificadores como enfoque fundamental y permite a los usuarios elegir y probar nuevos modelos, ofreciendo flexibilidad para la personalización (Walczyna & Piotrowski, 2023). DeepFaceLab, por su parte, incorpora avances como las redes generativas adversariales (GAN) junto con los autocodificadores. Estos algoritmos y soluciones de software proporcionan a los usuarios las herramientas necesarias e interfaces fáciles de usar para realizar intercambios de rostros de alta calidad (Walczyna & Piotrowski, 2023).

Además, DeepFaceLab ofrece una estructura de acoplamiento flexible y ligera, que permite a los usuarios integrar otras funciones en sus flujos de trabajo sin necesidad de codificación compleja y repetitiva (Perov et al., 2021).

Al utilizar estos algoritmos avanzados y soluciones de software, los investigadores y usuarios pueden explorar el potencial del intercambio de rostros en DeepFakes, garantizando al mismo tiempo la generación de resultados visualmente convincentes y realistas. El desarrollo y perfeccionamiento continuos de estas herramientas contribuyen significativamente al avance de la tecnología de DeepFakes amplía los límites de la personalización, el realismo y la facilidad de uso (Perov et al., 2021; Walczyna & Piotrowski, 2023).

Sincronización labial. Las deepfakes de sincronización labial, una forma de medios sintéticos, implican la modificación de un vídeo fuente para alinear los movimientos de la región bucal con una grabación de audio diferente (Agarwal et al., 2019). Esta técnica llamó la atención cuando el actor y director Jordan Peele manipuló un vídeo del presidente Obama para hacerle pronunciar frases como "El presidente Trump es un total y completo idiota" (Agarwal et al., 2019). Estos casos ponen de relieve el potencial de las deepfakes de sincronización labial para engañar a los espectadores creando la ilusión de que los individuos dicen cosas que en realidad nunca dijeron.

Según (Masood et al., 2021), uno de los principales objetivos de las falsificaciones de sincronización labial es mostrar a una persona hablando de la forma en que el atacante imagina que lo hace la víctima. Al sintetizar contenidos audiovisuales de gran realismo, las falsificaciones de sincronización labial tienen la capacidad de engañar a los espectadores y manipular la opinión pública. Las implicaciones de esta tecnología son profundas, ya que puede inducir a confusión, difundir información errónea e incluso dañar la reputación de personas específicas.

La capacidad de alterar el discurso y representar visualmente a personas que dicen declaraciones inventadas tiene importantes implicaciones éticas y sociales. En una época en la que los medios de comunicación desempeñan un papel crucial en la formación de la percepción pública, el auge de las falsificaciones labiales plantea dudas sobre la fiabilidad y autenticidad de los contenidos digitales. El posible uso indebido de esta tecnología puede tener graves consecuencias, como la difusión de información falsa, la difamación de figuras públicas y el debilitamiento de la confianza del público en las fuentes de los medios de comunicación.

Además, las falsificaciones de sincronización labial suponen una grave amenaza para la integridad de los medios digitales y la confianza de la sociedad. Se necesitan métodos de detección eficaces y medidas reguladoras exhaustivas para combatir las consecuencias

negativas asociadas a las suplantaciones de la sincronización labial y garantizar la autenticidad y fiabilidad de los contenidos digitales en el futuro.

Puppet Master. La técnica de deepfakes de marionetas es ampliamente reconocida y popular dentro del ámbito de los deepfakes (Gavrovska, 2022). En este enfoque, el atacante replica varios aspectos del comportamiento de la persona objetivo, incluidas las expresiones faciales, los movimientos oculares y los movimientos de la cabeza (Masood et al., 2021). El objetivo principal es integrar estos rasgos imitados, y a veces incluso el cuerpo entero, en un vídeo, animándolos según las preferencias del suplantador (Masood et al., 2021).

Los DeepFakes de marionetas han ganado considerable atención debido a su capacidad para producir resultados altamente realistas y persuasivos. Al imitar fielmente las señales de comportamiento de la persona objetivo, como las expresiones faciales y el lenguaje corporal, estas imitaciones pueden hacer creer a los espectadores que el imitador es una persona auténtica. La meticulosa atención al detalle y la precisión en la reproducción de estos matices físicos contribuyen a la autenticidad y credibilidad del contenido manipulado.

Las implicaciones de estos DeepFakes van más allá de la imitación visual, tienen el potencial de manipular la percepción pública, difundir información errónea y facilitar actividades fraudulentas. Al apropiarse de las expresiones y movimientos de la persona objetivo, un atacante puede crear vídeos que transmitan mensajes u opiniones que el individuo original nunca llegó a expresar. Esto suscita preocupación en cuanto a la fiabilidad e integridad de los medios digitales, ya que los espectadores pueden ser fácilmente engañados por estas sofisticadas manipulaciones.

Además, su uso plantea consideraciones éticas y jurídicas. La reproducción no autorizada de la imagen de una persona y su participación involuntaria en escenarios inventados infringe su derecho a controlar su propia imagen y su narrativa personal. Estas acciones pueden dañar la reputación de la persona, violar su intimidad e incluso causarle trastornos psicológicos.

Síntesis facial y manipulación de atributos. Los deepfakes de síntesis facial y manipulación de atributos generan imágenes fotorrealistas de rostros y editan atributos faciales. Esta manipulación se utiliza a menudo para difundir desinformación en las redes sociales a través de perfiles falsos (Masood et al., 2021).

Los métodos actuales de manipulación facial pueden clasificarse a grandes rasgos en las siguientes categorías: 1) manipulación de expresiones faciales, en la que se pueden transferir las expresiones faciales de una persona a otra utilizando un método como Face2Face, y 2) manipulación de la identidad basada en métodos de intercambio de caras, en la que se puede sustituir la cara de una persona por la de otra (Khalid & Woo, 2020).

Ilustración 3

Ejemplo de tipos de DeepFakes

Nota. Se exhiben cinco fotogramas ejemplares de un clip de 10 segundos, que muestran el material original, una sincronización labial generada mediante una imitación profunda, una imitación cómica, un intercambio de caras mediante una imitación profunda y un deepfake de maestro de marionetas (Agarwal et al., 2019).

Deepfakes de audio. Por último, los deepfakes de audio se centran en generar la voz del orador objetivo mediante técnicas de aprendizaje profundo, lo que permite representar al orador diciendo algo que en realidad no dijo. Las voces artificiales pueden crearse utilizando técnicas de síntesis de texto a voz (TTS) o de conversión de voz (VC) (Masood et al., 2021). La tarea de detectar deepfakes de audio implica la necesidad de buscar características de los datos que se correspondan con las etiquetas de verdad (auténtico frente a falsificado). Normalmente, cualquier anomalía descubierta se corresponde con la naturaleza de un deepfake, como un fallo de ruido, desajuste de fase, reverberación o pérdida de inteligibilidad (Müller et al., 2021).

Algoritmos Utilizados En La Creación De Deepfakes

Generative Adversial Networks (GAN). Las redes adversariales generativas (GAN) son mecanismos populares para sintetizar contenido. Un GAN enfrenta a dos redes neuronales, un generador y un discriminador, entre sí. Para sintetizar una imagen de una persona ficticia, el generador comienza con una matriz aleatoria de píxeles y aprende de forma iterativa a sintetizar una cara realista. En cada iteración, el discriminador aprende a distinguir la cara sintetizada de un corpus de caras reales; si la cara sintetizada es distinguible de las caras reales, entonces el discriminador penaliza al generador. A lo largo de múltiples iteraciones, el generador aprende a sintetizar caras cada vez más realistas hasta que el discriminador es incapaz de distinguirlo de las caras reales (Nightingale & Farid, 2022).

GAN combina dos redes neuronales, la generadora y la discriminadora, para crear imágenes realistas. La red generadora genera las imágenes falsas basándose en el conjunto de datos de imágenes proporcionado, con el objetivo de producir imágenes de aspecto auténtico. Por otro lado, la red discriminadora evalúa las imágenes producidas por el generador, valorando su autenticidad. Mediante el entrenamiento del generador y el discriminador utilizando el método mínimo-máximo, en el que el mínimo representa 0 y el

máximo representa 1, el discriminador ayuda a generar imágenes más realistas que parezcan auténticas a los observadores.

La red generativa adversarial requiere un entrenamiento suficiente para alcanzar su máximo potencial y generar resultados de alta calidad. El entrenamiento prolongado del generador y el discriminador conduce a la producción de imágenes DeepFake más realistas y auténticas, lo que permite transformar el rostro de la persona A en el de la persona B en un vídeo (Yadav & Salmani, 2019).

Otra técnica para hacer un DeepFake es utilizar dos conjuntos de codificadores-decodificadores con cargas divididas para la red de codificadores. Encontrar una manera de forzar ambas caras en el mismo codificador es lo que hace posible el DeepFake. Esto puede resolverse fácilmente haciendo que el mismo codificador comparta dos redes diferentes mientras utiliza simultáneamente dos decodificadores distintos. Así, el intercambio de caras se consigue cuando la cara de entrada se codifica y luego se decodifica utilizando el decodificador de la cara objetivo. Se necesitan dos conjuntos de imágenes de entrenamiento para entrenar el programa DeepFake. El primer conjunto consiste en imágenes de muestra de la cara que se va a sustituir. Estas muestras pueden obtenerse fácilmente de un vídeo. Para obtener resultados mejores y más realistas, el primer conjunto puede ampliarse con fotos de otras fuentes. El segundo conjunto consiste en fotos de la cara que se intercambiará en el vídeo. El proceso de entrenamiento de los autocodificadores se hace más rápido y más eficiente si los conjuntos de imágenes de los rostros objetivo y original tienen las mismas condiciones de iluminación con ángulos de visión similares. Si se consigue esto, el intercambio será más fácil y se podrán obtener mejores resultados. Sin embargo, si ambos autocodificadores se entrenan por separado, serán incompatibles entre sí y cada decodificador sólo podrá decodificar un único tipo de representación latente. Esto puede solucionarse forzando a los dos codificadores a compartir los pesos de la red de codificadores mientras se utilizan dos decodificadores diferentes.

Una vez completado el proceso de entrenamiento, la representación latente de la cara que se generó a partir del conjunto de entrenamiento se pasa a la red decodificadora

entrenada en el sujeto que se supone que se va a insertar en el vídeo. Una vez hecho esto, el decodificador intentará reconstruir una cara del nuevo sujeto a partir de la información relativa a la cara del sujeto original presente en el vídeo. El proceso se repite entonces para cada fotograma del vídeo en el que se requiera una operación de intercambio de caras (Albahar & Almalki, 2019).

Variational Autoencoders (VAE). Los autocodificadores se han utilizado ampliamente para generar imágenes desde el desarrollo de los autocodificadores variacionales (VAE). Un gran número de programas de manipulación facial conocidos se basan en autocodificadores, como DeepFakes y DeepFaceLab. Estos métodos tienden a aprender la información de identidad para la manipulación facial a través del proceso de reconstrucción. Sin embargo, suelen ajustarse al dominio específico y no pueden escalarse a múltiples identidades (Jiang et al., 2022).

Deep Neural Networks (DNNs). Los métodos de detección de Deepfake desarrollados recientemente se basan en redes neuronales profundas (DNN) para distinguir los vídeos falsos generados por IA de los vídeos reales (Hussain et al., 2021).

Convolutional Neural Networks (CNN). Los métodos más efectivos para detectar y prevenir deepfakes son enfoques basados en aprendizaje profundo, que utilizan redes neuronales convolucionales como base para una tarea de clasificación binaria. Estas redes neuronales convolucionales extraen los patrones subyacentes de los fotogramas de entrada y los transmiten a una red totalmente conectada encargada de clasificar dichos patrones como confiables o no confiables (Silva et al., 2022).

Tradicionalmente, las redes neuronales convolucionales (CNN) se han utilizado ampliamente para la detección de DeepFakes en vídeos, y los resultados más exitosos se han logrado mediante la implementación de métodos basados en EfficientNet B7 (Coccomini et al., 2022). Estos métodos han demostrado su eficacia a la hora de discernir entre contenido manipulado y auténtico.

Por ejemplo, según (Khalid & Woo, 2020) han empleado modelos de clasificación de imágenes basados en CNN para diferenciar las imágenes DeepFake de las auténticas. Aprovechando la potencia de las CNN, estos modelos pueden identificar y clasificar con precisión contenidos visuales manipulados, contribuyendo a los esfuerzos en curso en la detección de DeepFakes.

Las últimas técnicas de vanguardia para detectar contenido facial manipulado en vídeos se basan principalmente en redes neuronales convolucionales (CNN). Por lo general, un detector completo de DeepFakes comprende un mecanismo de rastreo facial, seguido de la transmisión de la región facial extraída a un clasificador basado en CNN para su clasificación como real o falsa (Hussain et al., 2021). Este enfoque ha mostrado resultados prometedores y se ha convertido en un componente esencial del proceso de detección de falsificaciones.

En los últimos años, se han desarrollado métodos basados en el aprendizaje profundo para identificar y detectar estas manipulaciones. Se han probado numerosas arquitecturas de CNN en conjuntos de datos supervisados para discriminar eficazmente entre imágenes generadas por redes generativas adversariales (GAN) e imágenes auténticas. Los resultados iniciales han demostrado ser prometedores a la hora de distinguir con precisión entre estos tipos de imágenes. Sin embargo, es crucial señalar que el rendimiento se degrada cuando existen discrepancias significativas entre los conjuntos de datos de entrenamiento y de prueba o cuando los datos se someten a técnicas de compresión (Montserrat et al., 2020).

Al ampliar la investigación y la comprensión de la detección de deepfakes basada en CNN, es posible desarrollar algoritmos más robustos y precisos para identificar contenidos manipulados, contribuyendo así a los esfuerzos en curso para combatir la propagación de la desinformación y proteger la integridad de los medios visuales.

Técnicas De Reconocimiento. Debido a su impresionante realismo, los vídeos de Deepfake presentan un desafío para el ojo humano, ya que resulta difícil discernir de manera directa la diferencia entre ellos y los vídeos auténticos. Según la literatura, no existe un método de detección automatizado y de alta precisión para identificar Deepfakes. La falta de disponibilidad de un método eficaz de detección de Deepfakes es un gran reto para el mundo debido a la facilidad con la que se generan vídeos Deepfake y a su rápida propagación. Sin embargo, hay muchos esfuerzos para resolver este fenómeno y los métodos relacionados con el aprendizaje profundo muestran un rendimiento notable que otros métodos (Weerawardana & Fernando, 2021).

Análisis Forense. Según (Agarwal et al., 2020), las técnicas para detectar vídeos DeepFake pueden clasificarse como enfoques forenses de bajo nivel o de alto nivel.

Las técnicas forenses de bajo nivel detectan artefactos a nivel de píxel introducidos por el proceso de síntesis. Algunas de estas técnicas detectan artefactos genéricos, mientras que otras detectan artefactos explícitos que resultan, por ejemplo, de la deformación de la imagen, la mezcla de imágenes y las incoherencias entre la imagen y los metadatos. Aunque estas técnicas detectan una variedad de falsificaciones con una precisión relativamente alta, el inconveniente es que pueden ser sensibles al blanqueo no intencionado (por ejemplo, transcodificación o redimensionamiento) o a ataques adversarios intencionados.

En contraste, los enfoques de alto nivel tienden a generalizarse y son más resistentes al blanqueo y a los ataques de adversarios. Estas técnicas se centran en características semánticamente significativas que incluyen, por ejemplo, incoherencias en los parpadeos, la postura de la cabeza, señales fisiológicas y gestos distintivos.

En el análisis forense de imágenes generadas por GAN reveló que los GAN dejan algunas huellas de alta frecuencia en las imágenes que generan (Montserrat et al., 2020).

Deepfake-Specific Methods. En una investigación preliminar llevada a cabo por investigadores, se analizaron diversas arquitecturas de redes neuronales convolucionales (CNN) en un entorno supervisado con el objetivo de distinguir entre imágenes generadas por GAN e imágenes reales. Se descubrió que varias de estas arquitecturas CNN resultaron altamente efectivas. No obstante, se observó una notable disminución en el rendimiento de estos modelos cuando los datos se comprimían utilizando el proceso estándar utilizado en las redes sociales. Además, cuando los niveles de compresión en los datos de entrenamiento y prueba no coincidían, se producía un deterioro aún mayor en el rendimiento. Asimismo, muchos de estos detectores mostraron una tendencia a sobre ajustarse al conjunto de entrenamiento, lo cual se tradujo en un rendimiento deficiente al enfrentarse a nuevos datos generados por modelos diferentes (Verdoliva, 2020).

Datasets Usados Por Herramientas De Reconocimiento Y Detección De Deepfakes

Deepfake Detection Challenge (DFDC). Con el objetivo de abordar esta creciente amenaza, se ha desarrollado un conjunto de datos extenso de vídeos de intercambio de caras para entrenar modelos de detección, como señala el trabajo de (Dolhansky et al., 2020). El Deepfake Detection Challenge (DFDC) fue anunciado en septiembre de 2019 como una iniciativa colaborativa entre la industria, la comunidad académica y las organizaciones de la sociedad civil, con el propósito de fomentar la investigación en detección de manipulación facial. Como parte integral de este desafío, se ha creado un conjunto de datos que consiste en una amplia colección de videos que presentan rostros humanos, junto con etiquetas descriptivas que indican si han sido generados mediante técnicas de manipulación facial. Es importante destacar que todos los videos del conjunto de datos han sido producidos en colaboración con actores remunerados, y se pondrá a disposición de la comunidad de manera gratuita para el desarrollo, prueba y análisis de técnicas de detección de videos con rostros manipulados (Dolhansky et al., 2019).

Además, en el marco del DFDC, se ha llevado a cabo el concurso Kaggle DeepFake Detection Challenge. Este concurso se ha destacado por emplear el conjunto de datos DFDC,

el cual se distingue como uno de los más extensos y accesibles públicamente en cuanto a videos de intercambio de caras. Consta de más de 100.000 clips obtenidos de la participación de 3.426 actores remunerados. Este conjunto de datos engloba una diversidad de métodos utilizados en la generación de DeepFakes, incluyendo aquellos basados en GAN y otros métodos no entrenados. Además de proporcionar una descripción detallada de la metodología empleada para construir el conjunto de datos, el equipo de investigación ofrece un análisis minucioso de las mejores presentaciones recibidas durante el concurso Kaggle, en el cual se evalúan las técnicas y enfoques más efectivos en la detección de DeepFakes (Dolhansky et al., 2020).

Este enfoque colaborativo y el acceso a un conjunto de datos tan extenso y diverso representan un hito significativo en el avance de la investigación en detección de DeepFakes. Al proporcionar un recurso valioso para la comunidad académica y los expertos en seguridad, el DFDC y el concurso Kaggle han promovido la innovación y el desarrollo de nuevas técnicas de detección que permitirán abordar de manera más efectiva el desafío que plantea la proliferación de manipulaciones faciales en los medios digitales.

Celeb-DF.

Ilustración 4

Comparativa de Fotogramas del Conjunto de Datos Celeb-DF

Nota. Fotogramas de Videos Reales en la Columna Izquierda y Fotogramas Deepfake en las Columnas Derechas.

El conjunto de datos Celeb-DF está compuesto por 590 vídeos reales y 5.639 vídeos DeepFake, lo que equivale a más de dos millones de fotogramas de vídeo. Cada vídeo tiene una duración promedio de alrededor de 13 segundos, con una frecuencia estándar de 30 fotogramas por segundo. Los vídeos reales fueron obtenidos de entrevistas públicas de 59 famosos en YouTube, abarcando una diversa distribución en términos de género, edad y grupo étnico.

La proporción de sujetos en los vídeos reales es del 56,8% de hombres y el 43,2% de mujeres. En cuanto a las edades, el 8,5% tiene 60 años o más, el 30,5% tiene entre 50 y 60 años, el 26,6% tiene 40 años, el 28,0% tiene 30 años y el 6,4% tiene menos de 30 años. En cuanto a la composición étnica, el 5,1% son asiáticos, el 6,8% afroamericanos y el 88,1% caucásicos.

Los vídeos reales exhiben una amplia variedad de cambios en diferentes aspectos, como el tamaño de la cara de los sujetos en píxeles, las orientaciones, las condiciones de iluminación y los fondos. Por otro lado, los vídeos DeepFake se generan mediante el intercambio de caras entre cada par de los 59 sujetos. Todos los vídeos finales están en formato MPEG4.0 (Li et al., 2020).

CelebA-HQ. El conjunto de datos de vídeos de famosos de alta calidad (CelebV-HQ) consta de 35.666 clips de vídeo con una resolución mínima de 512×512, en los que aparecen 15.653 identidades. Cada clip está etiquetado manualmente con 83 atributos faciales, que abarcan la apariencia, la acción y la emoción. Se realiza un análisis exhaustivo para demostrar la diversidad y la coherencia temporal del conjunto de datos. Este análisis abarca varios aspectos, como la edad, el origen étnico, la estabilidad del brillo, la suavidad del movimiento, la diversidad de poses de la cabeza y la calidad de los datos, y pone de relieve la gran variedad y coherencia del conjunto de datos (Zhu et al., 2022).

FaceForensics++. Es un conjunto de datos, compuesto por 1004 vídeos en los que aparecían distintas personas. La metodología de investigación consistió en recopilar y

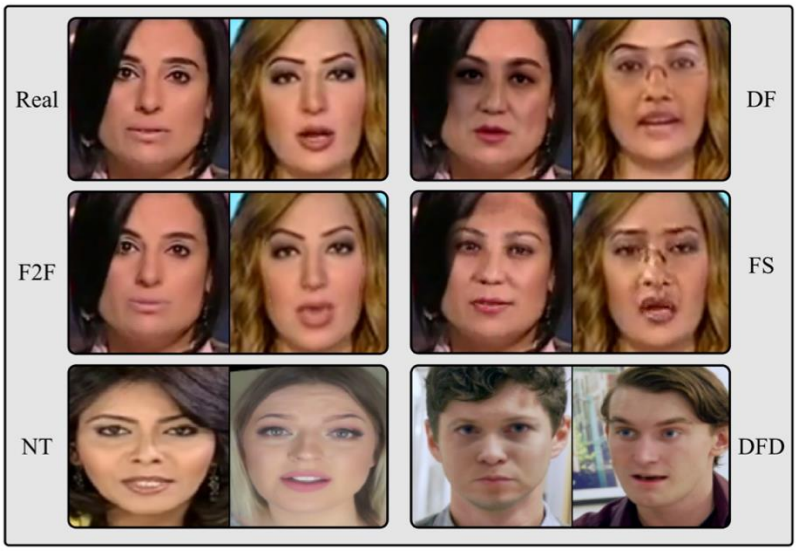
procesar datos para generar dos conjuntos de datos distintos. El primer conjunto de datos estaba formado por vídeos manipulados en los que los vídeos de origen y de destino eran diferentes, mientras que el segundo conjunto de datos estaba formado por vídeos creados mediante la tecnología Face2Face, que reproducían el vídeo de entrada con los mismos vídeos de origen y de destino. Este segundo conjunto de datos proporcionó pares de imágenes sintéticas y reales.

Los datos para el conjunto de datos se recogieron de YouTube, concretamente seleccionando vídeos con una resolución superior a 480p y etiquetas como "face" (cara), "newscaster" (presentador de noticias) o "newsprogram" (programa de noticias) del conjunto de datos youtube8m. También se descubrieron otros vídeos en YouTube utilizando estas etiquetas. Para identificar las secuencias de vídeo que contenían una presencia facial continua de más de 300 fotogramas, se utilizó el detector de caras de Viola-Jones. Se empleó un meticuloso proceso de selección manual para garantizar la selección de vídeos de alta calidad y evitar los vídeos con oclusiones faciales.

Para procesar los datos de vídeo se utilizó una versión modificada de la técnica Face2Face. Esta técnica facilitó la creación totalmente automática de manipulaciones de recreación al volver a representar el rostro en un vídeo de destino con expresiones variables. El preprocesamiento consistió en utilizar los fotogramas iniciales para obtener una identidad temporal del rostro (modelo 3D) y realizar un seguimiento de las expresiones a lo largo de los fotogramas restantes. Para mejorar el ajuste de la identidad y la textura estática, se empleó un enfoque automatizado para seleccionar los fotogramas con el ángulo izquierdo y derecho de la cara. Este paso, que originalmente requería intervención manual en la aplicación Face2Face, se simplificó. A partir de las poses seleccionadas, se reconstruyó la identidad del rostro y se realizó un seguimiento de todo el vídeo para calcular la expresión, la pose rígida y los parámetros de iluminación de cada fotograma (Rössler et al., 2018).

Ilustración 5

Los ejemplos incluyen imágenes reales y falsas de rostros humanos procedentes del conjunto de datos FaceForensics++.



Nota. En la primera fila se muestran imágenes auténticas, mientras que en las filas inferiores se muestran imágenes falsas procedentes de los conjuntos de datos DF, F2F, FS, NT y DFD (Khalid & Woo, 2020).

Herramientas De Reconocimiento Pagadas

Amber Authenticate. La herramienta conocida como Amber Authenticate funciona en segundo plano en un dispositivo mientras se realiza la captura de vídeo. Genera "hashes" a intervalos regulares, determinados por el usuario, que sirven como representaciones criptográficamente codificadas de los datos capturados. Estos hashes se registran de forma permanente en una cadena de bloques pública. Si se vuelve a procesar el mismo segmento de vídeo con el algoritmo, cualquier alteración de los datos de audio o vídeo del archivo producirá hashes diferentes. Esta discrepancia alerta al usuario de una posible manipulación o alteración (Newman, 2019).

Sensity. Fundada a finales de 2018, Sensity AI surgió como la empresa pionera mundial en inteligencia de amenazas visuales. Formada por un equipo dedicado de investigadores de aprendizaje automático y especialistas en inteligencia sobre amenazas, su misión principal es proteger a las personas y las organizaciones de los peligros que presentan los DeepFakes y otras formas de medios visuales malévolos (The World Economic Forum, 2023) .

Sensity, en su capacidad, posee la capacidad de discernir DeepFakes, incluidos los creados como identidades sintéticas mediante la utilización de redes generativas adversariales (GAN). En plataformas como las redes sociales y los sitios de citas, estos personajes fabricados y cuentas de bots son frecuentes. Los mecanismos de detección de Sensity han sido sometidos a un riguroso entrenamiento utilizando millones de imágenes generadas por GAN descubiertas en diversas plataformas en línea, que abarcan diversas condiciones de compresión, recorte y filtros fotográficos. Gracias a este entrenamiento, los detectores han adquirido la capacidad de identificar artefactos específicos y señales de alta frecuencia intrínsecas a las imágenes generadas por IA, características que suelen estar ausentes en las fotografías naturales (Sensity AI, 2023) .

Cogito Detect. Cogito es una de las empresas que ofrece tanto un servicio de detección de DeepFakes con un sistema de detección de DeepFakes de origen humano para todo tipo de imágenes o vídeos como expertos internos. Utilizando su enorme base de datos de este tipo de imágenes y vídeos, Cogito puede detectar DeepFake con un alto nivel de precisión. También proporciona conjuntos de datos de entrenamiento para la detección de DeepFakes con el fin de desarrollar un modelo de IA que pueda detectar los contenidos DeepFake con el mejor nivel de precisión a un precio asequible (Brown, 2020).

Deeptrace. Deeptrace, una empresa de ciberseguridad con sede en Ámsterdam, se especializa en la utilización de tecnologías de aprendizaje profundo y visión por ordenador para detectar y monitorizar medios sintéticos. Su objetivo principal es proteger a las personas y las organizaciones de los efectos perjudiciales de los medios sintéticos generados por IA.

Desde su creación en 2018, Deeptrace ha demostrado un compromiso inquebrantable con la investigación de las capacidades y amenazas en constante evolución asociadas con DeepFakes. A través de su investigación, proporcionan inteligencia esencial para mejorar su tecnología de detección. Deeptrace ha llevado a cabo recientemente un extenso análisis del panorama de las falsificaciones profundas, que ha dado como resultado su mapeo más completo hasta la fecha. Los datos compartidos en su informe arrojan luz sobre las consecuencias tangibles de los deepfakes en el mundo real.

Mediante la presentación de una visión general por parte de expertos, Deeptrace pretende cortar por lo sano las afirmaciones exageradas y la desinformación que rodea a los DeepFakes. Las conclusiones del informe se basan en datos de fuentes independientes y se complementan con las perspectivas de destacados expertos en la materia (Ajder et al., 2019).

Herramientas De Reconocimiento Open Source

FakeCatcher. FakeCatcher busca continuamente indicios auténticos dentro de vídeos auténticos, evaluando la esencia de nuestra humanidad: el matizado "flujo sanguíneo" presente en los píxeles de vídeo. A medida que nuestro corazón hace circular la sangre, nuestras venas experimentan variaciones de color. Estas señales de flujo sanguíneo se recogen de toda la región facial, y los algoritmos las convierten en mapas espaciotemporales. Posteriormente, empleando el aprendizaje profundo, podemos discernir rápidamente la autenticidad de un vídeo (Intel, 2022).

DeepFaceLab. DeepFaceLab es conocido como el principal software utilizado para crear DeepFakes, ya que más del 95% de los vídeos de DeepFakes se generan utilizando esta plataforma. La defensa contra las falsificaciones profundas exige investigar técnicas de detección, pero también avances en los métodos de generación. Sin embargo, los métodos de DeepFake actuales suelen adolecer de flujos de trabajo enrevesados y un rendimiento inferior.

Para hacer frente a este reto, DeepFaceLab surge como un marco de DeepFake dominante diseñado específicamente para el intercambio de rostros. No sólo proporciona herramientas esenciales, sino que también ofrece una interfaz fácil de usar para realizar intercambios de rostros de alta calidad sin problemas. Además, DeepFaceLab presenta una estructura flexible y adaptable, que permite a los usuarios incorporar funcionalidades adicionales a sus procesos sin necesidad de recurrir a complejos códigos repetitivos.

El informe profundiza en los principios subyacentes que guían la implementación de DeepFaceLab y ofrece una visión general de su pipeline. Los usuarios tienen la libertad de modificar cada aspecto del proceso sin esfuerzo para conseguir las personalizaciones que deseen. DeepFaceLab ha demostrado ser capaz de producir resultados de calidad cinematográfica con una fidelidad notable. La superioridad de este sistema se demuestra comparando su rendimiento con el de otros métodos de intercambio de rostros (DeepfakeVFX.com, 2021).

MesoNet. MesoNet es un método automatizado diseñado para detectar la manipulación facial en vídeos. Dado que las técnicas convencionales de análisis forense de imágenes no son adecuadas para el análisis de vídeos debido a la degradación de datos causada por la compresión, los investigadores adoptan un enfoque de aprendizaje profundo. Construyen dos redes con un número limitado de capas, dando prioridad al examen de las propiedades mesoscópicas dentro de las imágenes. Para evaluar la eficacia de estas redes eficientes, se realizan evaluaciones tanto en un conjunto de datos existente como en un conjunto de datos recopilado específicamente a partir de vídeos en línea. Las pruebas revelan una tasa de detección satisfactoria superior al 98% para DeepFakes (Afchar et al., 2018).

FWA. Face Warping Artifacts emplea una ResNet-50 para detectar vídeos DeepFake, con el objetivo de revelar los artefactos de deformación facial resultantes de las operaciones de redimensionamiento e interpolación en el algoritmo fundamental de generación de DeepFake. El entrenamiento de este modelo se lleva a cabo utilizando imágenes de caras recogidas por él mismo (Li et al., 2020).

MesoInception-4. MesoInception-4 es una red basada en CNN que se inspira en InceptionNet para identificar la manipulación de rostros en vídeos. La arquitectura de la red consta de dos módulos de incepción y dos capas de convolución tradicionales, intercaladas con capas de agrupamiento máximo. Tras estas capas, se emplean dos capas totalmente conectadas. En lugar de utilizar la pérdida de entropía cruzada convencional, los autores introducen el error cuadrático medio como métrica entre las etiquetas verdaderas y las predichas. Antes de introducir las imágenes en la red, se redimensionan a 256 × 256 dimensiones (Rössler et al., 2018).

Capsule. Utiliza estructuras de cápsula basadas en una red VGG19 como arquitectura principal para la clasificación de DeepFake. Este modelo se entrena con el conjunto de datos FaceForensics++ (H. H. Nguyen et al., 2019).

CAPÍTULO III

METODOLOGÍA DE LA INVESTIGACIÓN

La metodología adoptada en el presente trabajo de titulación se basa en una revisión sistemática, una metodología ampliamente reconocida y rigurosa para sintetizar y evaluar de manera exhaustiva las pruebas de investigación pertinentes. La génesis de las revisiones sistemáticas se fundamenta en la necesidad de tomar decisiones informadas que afecten a la vida de las personas, sustentadas en un entendimiento actualizado y completo de la evidencia de investigación relevante. Siguiendo esta premisa, esta revisión sistemática se esfuerza por recopilar y analizar de manera sistemática todas las pruebas empíricas que se ajusten a criterios de elegibilidad predefinidos, con el objetivo específico de abordar una pregunta de investigación específica. Al emplear métodos explícitos y meticulosos, seleccionados con el propósito de minimizar sesgos, se busca garantizar la confiabilidad y solidez de los hallazgos extraídos, permitiendo así la formulación de conclusiones fundamentadas para respaldar una toma de decisiones informada. La metodología de revisión sistemática, que fue pionera y meticulosamente desarrollada por (Higgins JPT et al., 2023), se establece como un marco altamente estructurado, transparente y reproducible. Esto involucra aspectos cruciales como la formulación previa de la pregunta de investigación, la delimitación clara del alcance de la revisión y los criterios de inclusión de los estudios, la búsqueda exhaustiva de investigaciones pertinentes, la consideración integral de posibles sesgos en los estudios incluidos y, finalmente, el análisis imparcial de las pruebas de investigación recopiladas para obtener conclusiones objetivas y bien fundamentadas.

Con el fin de alcanzar los objetivos propuestos, se ha desarrollado una ruta de trabajo que brinda claridad en los procesos y pasos a seguir, posibilitando la elaboración de una metodología sólida.

La ruta de trabajo proporciona una metodología académica sólida y bien fundamentada, brindando una guía confiable y sistemática para la realización de este trabajo de titulación, garantizando la validez y confiabilidad de los resultados obtenidos. Se han

seleccionado técnicas y herramientas apropiadas para el análisis de datos y la obtención de conclusiones significativas, teniendo en cuenta la naturaleza específica de los objetivos y el alcance del estudio.

Ilustración 6

Ruta de trabajo

Una vez adquirida una comprensión exhaustiva de los procesos de generación de DeepFakes y de la variedad de herramientas, técnicas, modelos y conjuntos de datos empleados para su detección, surge la necesidad de establecer métricas y puntos de referencia precisos. Estas métricas son fundamentales para facilitar el análisis comparativo y la clasificación de la eficacia de las distintas herramientas de detección de DeepFakes. Este capítulo tiene como objetivo evaluar dichas métricas a partir de la literatura existente escrita por otros investigadores.

Datasets

Para los propósitos de este trabajo y para delimitar el análisis, se dará principal consideración a los resultados obtenidos en los conjuntos de datos DFDC (DeepFake Detection Challenge), Celeb-DF y FaceForensics++. Estos tres conjuntos de datos serán los enfoques principales de evaluación para la detección de DeepFakes en la investigación, ya que son ampliamente reconocidos y utilizados en el campo. Otros conjuntos de datos podrían ser tomados en cuenta para obtener una visión más completa, pero estos tres serán los principales pilares de la evaluación en esta investigación.

El conjunto de datos DFDC ha sido empleado extensamente en la evaluación de diferentes algoritmos y enfoques de detección de DeepFakes, convirtiéndose en una referencia importante para este estudio.

Por otra parte, también se considerará el conjunto de datos Celeb-DF en el análisis. Diseñado específicamente para reducir la brecha entre los datos de entrenamiento de DeepFakes y los videos reales de celebridades manipulados, Celeb-DF proporciona un desafío adicional para los métodos de detección. Su enfoque en videos de alta calidad de famosos generados mediante un proceso de síntesis mejorado lo hace relevante para el estudio de las herramientas de detección más efectivas.

Centrarse en estos tres conjuntos de datos permitirá obtener una perspectiva sólida y completa sobre las herramientas de detección de DeepFakes que han demostrado un mejor rendimiento y precisión en un contexto más realista y desafiante.

En la Tabla 1 se muestra la cantidad de videos y fotogramas presentes en los tres conjuntos de datos DFDC, Celeb-DF y FaceForensics ++ (FF-DF).

Esta tabla proporcionará una visión general de la escala de cada conjunto de datos y permitirá comprender su magnitud para el análisis y evaluación en el trabajo.

Tabla 1

I*nformación básica de los conjuntos de datos que serán utilizados (Li et al., 2020).*

Dataset	# Real		# DeepFake		Release Date
	Video	Frame	Video	Frame	
DFDC	1,131	488.4k	4,113	1,783.3k	2019.10
Celeb-DF	590	225.4k	**5,639**	2,116.8k	2019.11
FF-DF	1,000	509.9k	1,000	509.9k	2019.01

Métricas

Antes de proceder a la comparación entre las diversas herramientas y modelos existentes, es crucial comprender las métricas utilizadas para evaluar la eficacia en la detección de DeepFakes en el contexto de este trabajo de titulación.

Según la investigación de (Lin et al., 2022), las métricas de evaluación más comúnmente empleadas en la detección de DeepFakes son el Área bajo la Curva (AUC), la Precisión (Precision) y la Exactitud (Accuracy). No obstante, otros investigadores, como (Tran et al., 2021), consideran que las métricas de evaluación ampliamente utilizadas también incluyen Recall, Precision y F1-score. La Precisión evalúa la exactitud de las identificaciones positivas, mientras que Recall mide la eficacia de detectar correctamente los casos positivos reales. El puntaje F1, al ser la media armónica de Precision y Recall, permite considerar ambas métricas simultáneamente al comparar distintos modelos de detección. La Exactitud, por su parte, cuantifica el porcentaje de predicciones correctas realizadas por el modelo.

Además, se emplea la curva ROC (Receiver Operating Characteristic) para representar el rendimiento de un modelo de clasificación, trazando dos parámetros clave: la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR). Con el fin de evaluar el rendimiento global de la clasificación, se calcula el Área bajo la Curva ROC (AUC). Esta medida agregada considera todas las clasificaciones posibles, proporcionando valiosa información sobre la capacidad del modelo para clasificar las predicciones, independientemente de sus valores absolutos. La invariabilidad del AUC a escala lo convierte en una herramienta útil para evaluar con eficacia las clasificaciones de las predicciones. Estas métricas serán fundamentales para el análisis y comparación de las herramientas y modelos de detección de DeepFakes en el transcurso de este capítulo.

Autores

A continuación, se revisarán diversas evaluaciones y benchmarks realizados por varios autores, destacando las herramientas más sobresalientes en el proceso de detección de DeepFakes. Uno de los trabajos relevantes en este ámbito es el presentado por (Li et al., 2020), quienes han desarrollado un conjunto de datos desafiante y a gran escala para el desarrollo y la evaluación de métodos de detección de DeepFakes.

El conjunto de datos Celeb-DF fue diseñado con el objetivo de reducir la discrepancia en calidad visual entre los datos utilizados para entrenar modelos de DeepFakes y los videos manipulados de celebridades que circulan en Internet. Este conjunto de datos contiene 5.639

videos de alta calidad de famosos generados mediante un proceso de síntesis mejorado, lo que lo hace idóneo para la evaluación del rendimiento de los métodos actuales de detección de DeepFakes.

A través del análisis realizado en base al conjunto de datos Celeb-DF, se evalúa el desempeño de estos métodos, permitiendo identificar aquellos enfoques que ofrecen resultados más efectivos en la detección de manipulaciones.

Se presenta el benchmark obtenido por los autores, utilizando el área bajo la curva (AUC) para comparar las herramientas de detección. Esta elección se justifica debido a que todos los métodos comparados analizan fotogramas individuales, obteniendo una puntuación de clasificación para cada uno. El uso de AUC a nivel de fotograma evita las diferencias causadas por distintos enfoques para agregar puntuaciones a nivel de fotograma en cada vídeo. Además, el AUC a nivel de fotograma evita la necesidad de calibrar los resultados de clasificación en diferentes conjuntos de datos, ya que mide la capacidad de discriminación del clasificador sin depender de un umbral de decisión específico.

Esta elección de métrica proporciona una evaluación más robusta y generalizable de los métodos de detección de DeepFakes, permitiendo una comparación más equitativa y objetiva de su desempeño. Al evitar sesgos causados por la agregación de puntuaciones y la calibración en diferentes conjuntos de datos, el AUC a nivel de fotograma proporciona una medida más consistente de la capacidad de detección de cada método, lo que garantiza resultados más confiables en el análisis del benchmark.

Tabla 2

Resumen de los métodos de detección de Deepfakes comparados (Li et al., 2020).

Methods	Model Type	Training Dataset	Repositories	Release Date
Two-stream [54]	GoogLeNet InceptionV3 [48]	SwapMe [54]	Unpublished code provided by the authors	2018.03
MesoNet [6]	Designed CNN	Unpublished	https://github.com/DariusAf/MesoNet	2018.09
HeadPose [53]	SVM	UADFV [53]	https://bitbucket.org/ericyang3721/headpose_forensic/	2018.11
FWA [28]	ResNet-50 [19]	Unpublished	https://github.com/danmohaha/CVPRW2019_Face_Artifacts	2018.11
VA-MLP [33]	Designed CNN	Unpublished	https://github.com/FalkoMatern/Exploiting-Visual-Artifacts	2019.01
VA-LogReg [33]	Logistic Regression Model			
Xception [40]	XceptionNet [12]	FaceForensics++ [40]	https://github.com/ondyari/FaceForensics	2019.01
Multi-task [34]	Designed CNN	FaceForensics [39]	https://github.com/nii-yamagishilab/ClassNSeg	2019.06
Capsule [36]	Designed CapsuleNet [42]	FaceForensics++	https://github.com/nii-yamagishilab/Capsule-Forensics-v2	2019.10
DSP-FWA	SPPNet [18]	Unpublished	https://github.com/danmohaha/DSP-FWA	2019.11

Tabla 3

Puntuaciones AUC a nivel de fotograma (%) de varios métodos en conjuntos de datos comparados. Valores marcados en negrita corresponden al mejor rendimiento (Li et al., 2020).

Methods↓ Datasets→	FF-DF [40]	DFDC [14]	**Celeb-DF**
Two-stream [54]	70.1	61.4	53.8
Meso4 [6]	84.7	75.3	54.8
MesoInception4	83.0	73.2	53.6
HeadPose [53]	47.3	55.9	54.6
FWA [28]	80.1	72.7	56.9
VA-MLP [33]	66.4	61.9	55.0
VA-LogReg	78.0	66.2	55.1
Xception-raw [40]	**99.7**	49.9	48.2
Xception-c23	99.7	72.2	65.3
Xception-c40	95.5	69.7	**65.5**
Multi-task [34]	76.3	53.6	54.3
Capsule [36]	96.6	53.3	57.5
DSP-FWA	93.0	**75.5**	64.6

Por otro lado, el estudio realizado por (Tran et al., 2021) tiene como objetivo presentar un modelo de alto rendimiento que garantice la precisión y aborde el problema del peso excesivo asociado al uso de extensas redes dorsales (backbones) en los enfoques actuales.

Para lograr este propósito, proponen una metodología integral que incorpora varias técnicas innovadoras. El modelo propuesto incluye la extracción manual por destilación y de regiones específicas del objetivo, además del aumento de datos y el conjunto de varias regiones. Todo ello se complementa con una arquitectura basada en una red neuronal convolucional (CNN) que ofrece una clasificación flexible utilizando un umbral dinámico. Estas estrategias están diseñadas para reducir el problema del sobreajuste, el cual afecta la calidad y generalización de muchos modelos de detección de Deepfakes.

Para evaluar la eficacia del modelo propuesto, realizaron pruebas exhaustivas en dos conjuntos de datos: el conjunto de datos de detección de DeepFakes (DFDC) y Celeb-DF v2. Además, llevaron a cabo el entrenamiento y evaluación de los conjuntos de datos utilizando la tarjeta gráfica NVIDIA GeForce RTX 2080 Ti.

A continuación, se presentan los resultados obtenidos:

Tabla 4

Comparación del rendimiento utilizando el valor AUC (Tran et al., 2021).

	Celeb-DF v2	DFDC	Number of Parameters
Zhou et al. (2017) [5]	53.8%	61.4%	24 M
Afchar et al. (2018) [6]	54.8%	75.3%	27.9 K
Rossler et al. (2019) [47]	48.2%	49.9%	22.8 M
Nguyen et al. (2019) [53]	57.5%	53.3%	3.9 M
Li et al. (2020) [54]	64.6%	75.5%	-
Chen et al. (2021) [28]	90.56%	-	42.8 K
Heo et-al. (2021) [4]	-	97.8%	-
Ours	**97.8%**	**95.8%**	**26 M**

Tabla 5

Resultados del Benchmark del modelo (Tran et al., 2021).

	Precision	Recall	F1-Score	Accuracy
Ours–Celeb-DF v2	0.9825	0.9438	0.9628	0.9749
Ours–DFDC	0.9254	0.9232	0.9243	0.9244

Tabla 6

Resultados de la precisión de la red clasificadora en el conjunto de validación de Celeb-DF v2 y DFDC (Tran et al., 2021).

	Accuracy of Validation Set	Loss of Validation Set
Celeb-DF v2	94.2%	0.3
DFDC	93.75%	0.295

Al igual que (Tran et al., 2021), (Yan et al., 2023) proponen mejoras para la comparación y evaluación de las herramientas existentes para detectar DeepFakes. Ambos estudios reconocen el desafío crítico y a menudo subestimado en el ámbito de la detección de DeepFakes debido a la falta de un punto de referencia estandarizado y completo. Esta carencia genera comparaciones injustas y resultados potencialmente engañosos, ya que los modelos de detección se enfrentan a entradas incoherentes debido a la falta de uniformidad en los canales de procesamiento de datos. Además, las variaciones en los ajustes

experimentales y la falta de estrategias y métricas de evaluación estandarizadas agravan el problema.

En respuesta a esta preocupación, un equipo de investigación liderado por (Yan et al., 2023) presenta una solución innovadora llamada DeepfakeBench, que constituye el primer punto de referencia completo para la detección de DeepFakes. Este es pionero en tres aspectos fundamentales. En primer lugar, se establece un sistema unificado de gestión de datos que asegura entradas coherentes en todos los detectores. En segundo lugar, se integra un marco que permite aplicar los métodos más avanzados de detección de falsificaciones profundas. Por último, se establecen métricas y protocolos de evaluación estandarizados que fomentan la transparencia y la reproducibilidad en la investigación.

DeepfakeBench se destaca por su base de código extensible y modular, que incluye 15 métodos de detección de última generación, 9 conjuntos de datos de DeepFakes y una serie de protocolos de evaluación y herramientas de análisis. Además, el equipo de investigación proporciona análisis exhaustivos y valiosos de estas evaluaciones desde diversas perspectivas, como la exploración de aumentos de datos y backbones utilizados.

Tabla 7

Resumen de las herramientas de detección de Deepfakes comparadas (Yan et al., 2023).

Model Type	Detectors	Backbone	Repositories	Reference
Naive Detector	MesoNet [1]	Designed CNN	https://github.com/DariusAf/MesoNet	WIFS-2018
Naive Detector	MesoInception [1]	Designed CNN	https://github.com/DariusAf/MesoNet	WIFS-2018
Naive Detector	CNN-Aug [38]	ResNet [12]	https://peterwang512.github.io/CNNDetection/	CVPR-2020
Naive Detector	EfficientNet-B4 [31]	EfficientNet [31]	https://github.com/lukemelas/EfficientNet-PyTorch	ICML-2019
Naive Detector	Xception [27]	Xception [4]	https://github.com/ondyari/FaceForensics	ICCV-2019
Spatial Detector	Capsule [23]	Designed Capsule [28]	https://github.com/nii-yamagishilab/Capsule-Forensics-v2	ICASSP-2019
Spatial Detector	DSP-FWA [17]	Xception [4]	https://github.com/danmohaha/CVPRW2019_Face_Artifacts	CVPRW-2019
Spatial Detector	Face X-ray [15]	HRNet [37]	Unpublished code provided by the authors	CVPR-2020
Spatial Detector	FFD [5]	Xception [4]	cvlab.cse.msu.edu/project-ffd.html	CVPR-2020
Spatial Detector	CORE [24]	Xception [4]	https://github.com/niyunsheng/CORE	CVPRW-2022
Spatial Detector	RECCE [2]	Designed Networks	https://github.com/VISION-SJTU/RECCE	CVPR-2022
Spatial Detector	UCF [40]	Xception [4]	Unpublished code provided by the authors	ArXiv-2023
Frequency Detector	F3Net [26]	Xception [4]	Unpublished code provided by the authors	ECCV-2020
Frequency Detector	SPSL [20]	Xception [4]	Unpublished code provided by the authors	CVPR-2021
Frequency Detector	SRM [21]	Xception [4]	Unpublished code provided by the authors	CVPR-2021

El modelo propuesto por los autores incorpora una amplia colección de nueve conjuntos de datos reconocidos y ampliamente utilizados en el campo de la detección de deepfakes. Estos conjuntos de datos incluyen FaceForensics++ (FF++), CelebDF-v1

(CDFv1), CelebDF-v2 (CDFv2), DeepFakeDetection (DFD), DeepFake Detection Challenge Preview (DFDC-P), DeepFake Detection Challenge (DFDC), UADFV, FaceShifter (Fsh) y DeeperForensics-1.0 (DF-1.0). Entre ellos, el conjunto de datos FF++ contiene cuatro tipos de métodos de manipulación: Deepfakes (FF-DF), Face2Face (FF-F2F), FaceSwap (FF-FS) y NeuralTextures (FF-NT). Además, el estudio utiliza quince herramientas de detección mencionadas en la Tabla 7.

En la evaluación, (Yan et al., 2023) emplearon cuatro métricas ampliamente utilizadas: Accuracy (ACC), el área bajo la curva ROC (AUC), Average Precision (AP) y Equal Error Rate o la tasa de error igual (EER). Estas métricas permiten evaluar el rendimiento y la eficacia de las herramientas de detección en los distintos conjuntos de datos, proporcionando una visión integral de su capacidad para detectar deepfakes.

Con el uso de estos conjuntos de datos variados y herramientas de detección relevantes, el modelo propuesto brinda una evaluación exhaustiva y confiable de los métodos de detección de deepfakes, contribuyendo significativamente al avance de la investigación en este campo crucial.

Tabla 8

Evaluaciones dentro del dominio y entre dominios utilizando la métrica AUC (Yan et al., 2023).

Type	Detector	Backbone	Within Domain Evaluation								Cross Domain Evaluation									
			FF++_c23	FF++_c40	FF-DF	FF-F2F	FF-FS	FF-NT	Avg.	Top3	CDFv1	CDFv2	DF-1.0	DFD	DFDC	DFDCP	Fsh	UADFV	Avg.	Top3
Naive	Meso4 [1]	MesoNet	0.6077	0.5920	0.6771	0.6170	0.5946	0.5701	0.6097	0	0.7358	0.6091	0.9113	0.5481	0.5560	0.5994	0.5660	0.7150	0.6551	1
Naive	MesoIncep [1]	MesoNet	0.7583	0.7278	0.8542	0.8087	0.7421	0.6517	0.7571	0	0.7366	0.6966	0.9233	0.6069	0.6226	0.7561	0.6438	0.9049	0.7364	3
Naive	CNN-Aug [38]	ResNet	0.8493	0.7846	0.9048	0.8788	0.9026	0.7313	0.8419	0	0.7420	0.7027	0.7993	0.6464	0.6361	0.6170	0.5985	0.8739	0.7020	0
Naive	Xception [27]	Xception	0.9637	0.8261	0.9799	0.9785	0.9833	0.9385	0.9450	4	0.7794	0.7365	0.8341	**0.8163**	0.7077	0.7374	0.6249	0.9379	0.7718	2
Naive	EfficientB4 [31]	Efficient	0.9567	0.8150	0.9757	0.9758	0.9797	0.9308	0.9389	0	0.7909	0.7487	0.8330	0.8148	0.6955	0.7283	0.6162	0.9472	0.7718	3
Spatial	Capsule [23]	Capsule	0.8421	0.7040	0.8669	0.8634	0.8734	0.7804	0.8217	0	0.7909	0.7472	0.9107	0.6841	0.6465	0.6568	0.6465	0.9078	0.7488	2
Spatial	FWA [17]	Xception	0.8765	0.7357	0.9210	0.9000	0.8843	0.8120	0.8549	0	0.7897	0.6680	**0.9334**	0.7403	0.6132	0.6375	0.5551	0.8539	0.7239	1
Spatial	X-ray [15]	HRNet	0.9592	0.7925	0.9794	**0.9872**	0.9871	0.9290	0.9391	3	0.7093	0.6786	0.5531	0.7655	0.6326	0.6942	**0.6553**	0.8989	0.6985	0
Spatial	FFD [5]	Xception	0.9624	0.8237	0.9803	0.9784	0.9853	0.9306	0.9434	1	0.7840	0.7435	0.8609	0.8024	0.7029	0.7426	0.6056	0.9450	0.7733	1
Spatial	CORE [24]	Xception	0.9638	0.8194	0.9787	0.9803	0.9823	0.9339	0.9431	2	0.7798	0.7428	0.8475	0.8018	0.7049	0.7341	0.6032	0.9412	0.7694	0
Spatial	Recce [2]	Designed	0.9621	0.8190	0.9797	0.9779	0.9785	0.9357	0.9422	1	0.7677	0.7319	0.7985	0.8119	0.7133	0.7419	0.6095	0.9446	0.7649	2
Spatial	UCF [40]	Xception	**0.9705**	**0.8399**	**0.9883**	0.9840	**0.9896**	**0.9441**	**0.9527**	**6**	0.7793	0.7527	0.8241	0.8074	**0.7191**	**0.7594**	0.6462	**0.9528**	0.7801	**5**
Frequency	F3Net [26]	Xception	0.9635	0.8271	0.9793	0.9796	0.9844	0.9354	0.9449	1	0.7769	0.7352	0.8431	0.7975	0.7021	0.7354	0.5914	0.9347	0.7645	0
Frequency	SPSL [20]	Xception	0.9610	0.8174	0.9781	0.9754	0.9829	0.9299	0.9408	0	**0.8150**	**0.7650**	0.8767	0.8122	0.7040	0.7408	0.6437	0.9424	**0.7875**	3
Frequency	SRM [21]	Xception	0.9576	0.8114	0.9733	0.9696	0.9740	0.9295	0.9359	0	0.7926	0.7552	0.8638	0.8120	0.6995	0.7408	0.6014	0.9427	0.7760	2

En la Tabla 8, se presenta la evaluación realizada tanto dentro del dominio como entre dominios, así como la evaluación de manipulación cruzada. El objetivo de la evaluación entre dominios es analizar el rendimiento del modelo dentro del mismo conjunto de datos, mientras que la evaluación entre dominios consiste en probar el modelo en diferentes conjuntos de datos. También se llevó a cabo una evaluación de manipulación cruzada para medir el desempeño del modelo frente a distintos tipos de falsificaciones dentro del mismo conjunto de datos.

En esta tabla, se muestra la media del área bajo la curva (AUC) para las evaluaciones dentro del dominio y entre dominios, junto con los resultados globales. Además, se indica el número de métodos clasificados entre los tres primeros en todos los conjuntos de datos de prueba, lo cual se denota como "Top3". En cada columna, el método con los mejores resultados es resaltado en rojo, destacando así las herramientas más efectivas en cada caso.

Por otro lado, (Saikia et al., 2022) presentan un nuevo método para identificar con precisión la autenticidad de los vídeos mediante una técnica de extracción de características basada en el flujo óptico. El método extrae características temporales de los vídeos, que luego se introducen en un modelo híbrido que combina arquitecturas CNN y redes neuronales recurrentes (RNN) para la clasificación.

La investigación se centra en un enfoque híbrido de aprendizaje profundo que modela tanto las características dentro del fotograma como entre fotogramas de los vídeos para detectar eficazmente el contenido DeepFake. Para complementar este enfoque, se incorpora un método tradicional de análisis de características temporales, a saber, el flujo óptico, para ayudar en la extracción de características temporales. La implementación del flujo óptico caracteriza el movimiento de la cara del sujeto, explotando las posibles disimilitudes entre fotogramas.

Los autores han aplicado la metodología propuesta a tres conjuntos de datos: FaceForensics++, Celeb-DF y Deep Fake Detection Challenge (DFDC). Se realiza un análisis exhaustivo del rendimiento del modelo propuesto utilizando métricas clave como Accuracy, Recall, Precision, F1-score y AUC.

En particular, el estudio se centra en los rasgos faciales de los vídeos, ya que uno de los rasgos más comunes de los medios manipulados es la sustitución del perfil de un individuo por el de otra persona. Al centrarse en las características faciales, el modelo puede detectar artefactos distinguibles resultantes de la deformación en vídeos DeepFake.

Tabla 9

Comparación del rendimiento de los distintos modelos base (Saikia et al., 2022).

	DFDC (Frames:20)			Celeb-DF (Frames:50)			FF++ (Frames:30)		
Batch Size:128	**Test Accuracy**	**F1**	**AUC**	**Test Accuracy**	**F1**	**AUC**	**Test Accuracy**	**F1**	**AUC**
VGG16	**64.73%**	**64.28%**	**0.64**	67.09%	68.14%	0.68	78.39%	78.45%	0.78
InceptionV3	45.19%	50.00%	0.5	55.12%	52.27%	0.52	51.00%	50.51%	0.5
ResNet50	64.58%	59.87%	0.59	67.09%	68.44%	0.68	**89.67%**	**89.65%**	**0.89**
Xception	63.20%	64.19%	0.64	59.22%	63.49%	0.63	73.86%	73.04%	0.73
MobileNetV2	57.09%	55.68%	0.55	63.24%	65.21%	0.65	76.63%	76.76%	0.76
EfficientNetB7	59.84%	56.16%	0.56	**70.08%**	**69.13%**	**0.69**	83.66%	84.04%	0.84

Tabla 10

Comparación del rendimiento de los modelos básicos con el flujo óptico como elemento de partida en varios conjuntos de datos (Saikia et al., 2022).

	Celeb DF				DFDC				FF++			
	Accuracy	**Precision**	**Recall**	**AUC**	**Accuracy**	**Precision**	**Recall**	**AUC**	**Accuracy**	**Precision**	**Recall**	**AUC**
OF+RNN	52.13%	26.06%	34.21%	0.5	54.08%	24.96%	35.08%	0.5	47.73%	23.68%	50%	0.5
OF+CNN	**83.33%**	**83.78%**	**83.71%**	**0.83**	**69.77%**	**69.36%**	**68.64%**	**0.68**	89.19%	89.51%	88.92%	0.88
OF+RNN+CNN	79.49%	82.49%	79.08%	0.79	66.26%	67.11%	65.73%	0.66	**91.21%**	**91.20%**	**91.21%**	**0.91**

Tabla 11

Comparación del rendimiento del modelo híbrido en varios conjuntos de datos en función del número de fotogramas (Saikia et al., 2022).

	DFDC Dataset			FF++ Dataset			CelebDF Dataset		
Frames	**Accuracy**	**Precision**	**AUC Score**	**Accuracy**	**Precision**	**AUC Score**	**Accuracy**	**Precision**	**AUC Score**
10	64.58%	64.39%	0.63	74.87%	75.71%	0.75	63.25%	66.67%	0.65
20	64.27%	65.99%	0.64	78.39%	78.80%	0.78	67.09%	68.64%	0.68
30	**66.26%**	67.11%	**0.66**	83.17%	83.26%	0.83	73.07%	72.94%	0.73
40	64.12%	**69.04%**	0.63	77.89%	82.31%	0.79	73.08%	72.84%	0.73
50	-	-	-	86.68%	86.95%	0.87	78.21%	78.19%	0.78
60	-	-	-	83.17%	86.32%	0.84	72.65%	76.33%	0.73
70	-	-	-	**91.21%**	**91.20%**	**0.91**	74.36%	76.34%	0.74
80	-	-	-	-	-	-	76.07%	81.83%	0.76
90	-	-	-	-	-	-	**79.49%**	**82.69%**	**0.79**
100	-	-	-	-	-	-	76.07%	79.92%	0.76

Tabla 12

Comparación del modelo propuesto con los trabajos existentes (Saikia et al., 2022).

Paper	Method	Model	Datasets	Accuracy	AUC
Irene Amerini, et. al. [20]	Optical flow + CNN	VGG16	FF++	81.61%	-
		ResNet 50	FF++	75.46%	-
Peng Chen, et. al. [18]	CNN & LSTM	VGG16	FF++	-	100
			Celeb-DF	-	77.6
			UADFV	-	91.1
			Deepfake TIMIT HQ	-	98.5
			Deepfake TIMIT LQ	-	99.5
D Afchar, et al. [6]	CNN	Meso-4,	FF++ (Face2face)	95%	-
			Deepfake		-
		Meso-Inception-4	FF++ (Face2face)	98%	-
			Deepfake		-
Shivangi Aneja, et al. [21]	CNN & LSTM	Resnet18	FF++	92.23%	-
			Google DFD	81.21%	-
			AIF	60.79%	-
			Dessa	74.28%	-
			Celeb-DF	68.83%	-
			Combined	75.47%	-
Pranjal Ranjan, et. al [37]	CNN + LSTM	Xception Net	Celeb-DF	83.49%	-
			DFDC	78.13%	-
			DFD	94.33%	-
			Combined	79.62%	-
X Li, et. al [38]	Multiple Instance Learning	Xception Net	Celeb-DF	85.11%	-
			DFDC	98.84%	-
			FFPMS	90.71%	-
De Lima, et. al [39]	Spatio-temporal Convolutional Networks	RCN	Celeb-DF	76.25%	74.87
		R2Plus1D		98.07%	99.43
		I3D		92.28%	97.59
		R3D		98.26%	99.73
		MC3		97.49%	99.30
SA Khan, et. al [40]	CNN	VGG16	DFDC	96.75%	-
Our Work	Optical Flow + CNN + LSTM	VGG16	FF++	91.21%	0.91
			Celeb-DF	79.49%	0.79
			DFDC	66.26%	0.66

Este trabajo se basa en la utilización de vectores de flujo óptico con un modelo CNN pre entrenado, complementado con capas LSTM para capturar el movimiento irregular de cada píxel dentro de los fotogramas de vídeo. El modelo resultante puede evaluar y clasificar los vídeos como falsos o auténticos. Para hacer frente a las limitaciones computacionales, el experimento se realizó con un subconjunto de fotogramas en lugar de utilizar todos los fotogramas de los vídeos, ya que esto último exigiría unos recursos computacionales significativamente mayores. No obstante, los experimentos demostraron que el rendimiento del modelo mejoraba al aumentar el número de fotogramas por vídeo.

(Tolosana et al., 2020), presentan un análisis minucioso de la primera y la segunda generación de DeepFakes, centrándose en las regiones faciales y en el rendimiento de la detección de imitaciones. La investigación explora dos métodos distintos dentro de su marco experimental. El primer método sigue el enfoque tradicional predominante en la bibliografía,

que consiste en seleccionar el rostro completo como entrada para el sistema de detección de falsificaciones. El segundo método, sin embargo, introduce un enfoque novedoso en el que se eligen regiones faciales específicas como entrada para el sistema de detección de falsificaciones.

Tabla 13

Resultados de la detección de falsificaciones en términos de AUC (%)

AUC (%)	
Xception	
FaceForensics++ (2019) [15]	**99.40**
Celeb-DF (2019) [6]	**83.60**
DFDC Preview (2019) [7]	**91.17**
Capsule Network	
FaceForensics++ (2019) [15]	**99.52**
Celeb-DF (2019) [6]	**82.46**
DFDC Preview (2019) [7]	**87.45**

CAPÍTULO IV

DESARROLLO DEL PROYECTO

Con el propósito de evaluar el rendimiento de las herramientas, se llevó a cabo un análisis exhaustivo de la literatura existente, con el objetivo de determinar los parámetros para el benchmark propuesto en este proyecto de titulación. En primer lugar, se identificó que los estudios emplean el Área bajo la Curva (AUC) como métrica de evaluación, junto con otras métricas como Precisión, Recall, Accuracy y F1-score. Asimismo, se encontraron otras métricas relevantes, tales como dos variantes de Intersection over Union (IoU) y L1 distance , así como algunas variaciones de las métricas previamente mencionadas, como Average Precision (AP) y Equal Error Rate (EER) (Yan et al., 2023).

No obstante, se ha optado por utilizar exclusivamente el Área bajo la Curva (AUC) como métrica de evaluación de las herramientas de reconocimiento de DeepFakes debido a su sensibilidad para discriminar entre DeepFakes y videos auténticos, independencia del umbral de decisión, evaluación a nivel de fotograma que asegura una detección precisa incluso en casos sutiles de manipulación, y su capacidad para proporcionar una evaluación robusta y generalizable en diferentes conjuntos de datos y escenarios.

Esta elección meticulosa de utilizar el Área bajo la Curva (AUC) como métrica principal se justifica por su relevancia en el campo del reconocimiento de DeepFakes y su potencial para contribuir al avance de las técnicas de detección en la actualidad. Además, los resultados obtenidos con esta métrica proporcionan una sólida base para futuras investigaciones y desarrollos en este campo en constante evolución.

El enfoque en la utilización del AUC como métrica de evaluación se alinea con el propósito de este trabajo de tesis, el cual busca contribuir al mejoramiento de la detección de DeepFakes y a la protección de la integridad de la información frente a los retos planteados por el creciente uso de la inteligencia artificial y tecnologías afines.

En lo concerniente a los conjuntos de datos utilizados, se seleccionaron FaceForensics++, Celeb-DF y DFDC, debido a que abarcan una amplia gama de escenarios

de generación y manipulación de DeepFakes, lo cual posibilita una evaluación más completa y sólida de las técnicas de detección que se quieren evaluar.

Para llevar a cabo la comparación de las mejores herramientas, se utilizará el Área bajo la Curva (AUC) como métrica de evaluación, además de los mismos conjuntos de datos empleados en los estudios mencionados en el Capítulo III. Esta decisión ha sido tomada con el propósito de asegurar una base de comparación consistente y justa, lo que permitirá obtener resultados confiables y significativos en el análisis de rendimiento.

Además, al utilizar las mismas métricas que los estudios previos, se busca facilitar la comparación directa y objetiva de los resultados obtenidos. Al estandarizar las métricas de evaluación, se evita la introducción de posibles sesgos y se establece una base sólida para la valoración imparcial del rendimiento de las herramientas de detección de DeepFakes.

A continuación, se puede observar en la Tabla 13, las herramientas utilizadas en cada uno de los estudios realizados por los autores anteriormente. Se encuentran sombreados del mismo color las herramientas que serán comparadas directamente para que los resultados sean lo más precisos posible.

Tabla 14

Herramientas de detección de DeepFakes.

	(Li et al., 2020)	(Yan et al., 2023)	(Saikia et al., 2022)	(Tolosana et al., 2020)
Herramientas	Meso4	Meso4	-	-
	FWA	FWA	-	-
	MesoInception4	MesoInception4	-	-
	-	Xception	Xception	Xception
	Capsule	Capsule	-	Capsule
	HeadPose	EfficientB4	-	-
	VA-MLP	CNN-Aug	-	-
	VA-LogReg	X-ray	-	-
	Xception-raw	FFD	VGG16	-
	Xception-c23	CORE	InceptionV3	-
	Xception-c40	Reece	ResNet50	-
	Multi-task	UCF	MobileNetV2	-
	DSP-FWA	F3Net	EfficientNetB7	-
	Two-stream	SPSL	-	-
	-	SRM	-	-

Se llevará a cabo una comparación entre los resultados de las herramientas Meso4, FWA, MesoInception4, Xception y Capsule de los autores mencionados en el Capítulo III. Los resultados de estas herramientas serán evaluados teniendo en cuenta exclusivamente la métrica de AUC (Área bajo la curva), con el objetivo de analizar el rendimiento de cada.

Tabla 15

Comparación de la herramienta Meso4 utilizando la métrica AUC (%).

Herramienta	Autor	AUC (%)		
		DFDC	FaceForensics++	Celeb-DF
Meso4	Li et al., 2020	**75.30%**	**84.70%**	54.80%
	Yan et al., 2023	55.60%	60.8%	**73.58%**

Tabla 16

Comparación de la herramienta MesoInception4 utilizando la métrica AUC (%).

Herramienta	Autor	AUC (%)		
		DFDC	FaceForensics++	Celeb-DF
MesoInception4	Li et al., 2020	**73.20%**	**83.00%**	53.60%
	Yan et al., 2023	62.26%	75.83%	**73.66%**

Tabla 17

Comparación de la herramienta FWA utilizando la métrica AUC (%).

Herramienta	Autor	AUC (%)		
		DFDC	FaceForensics++	Celeb-DF
FWA	Li et al., 2020	**72.70%**	80.10%	56.90%
	Yan et al., 2023	61.32%	**87.65%**	**78.97%**

Tabla 18

Comparación de la herramienta Capsule utilizando la métrica AUC (%).

Herramienta	Autor	AUC (%)		
		DFDC	FaceForensics++	Celeb-DF
Capsule	Li et al., 2020	53.30%	96.60%	57.5%
	Yan et al., 2023	64.65%	84.21%	79.1%
	Tolosana et al., 2020	**87.40%**	**99.50%**	**82.4%**

Tabla 19

Comparación de la herramienta Xception utilizando la métrica AUC (%).

Herramienta	Autor	AUC (%)		
		DFDC	FaceForensics++	Celeb-DF
Xception	Yan et al., 2023	70.77%	9.37%	77.94%
	Saikia et al., 2022	64.00%	73.00%	63.00%
	Tolosana et al., 2020	**91.10%**	**99.50%**	**83.60%**

Ilustración 7

Comparación del rendimiento de herramientas de detección de DeepFakes según el conjunto de datos utilizado.

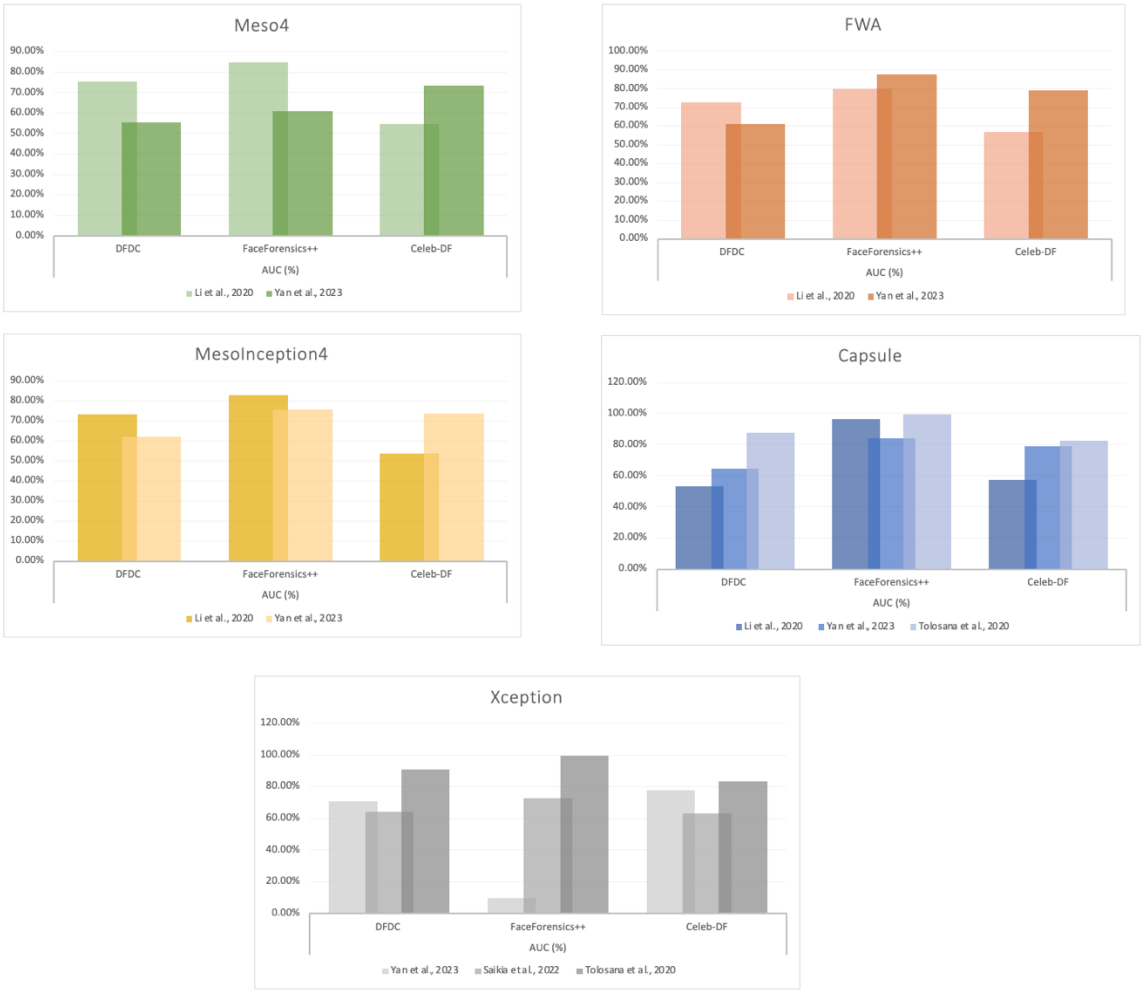


Tabla 20

Puntuación media de los resultados de la métrica AUC (%) en distintos conjuntos de datos.

Herramienta	*Promedio de AUC*
Meso4	67.46%
MesoInception4	70.26%
FWA	72.94%
Capsule	**78.30%**
Xception	70.25%

La herramienta Capsule fue la que mejor porcentaje promedio de AUC tuvo en comparación a las otras herramientas, cabe destacar que este resultado fue tomado en base a la comparación de tres autores a diferencia del segundo puntaje más alto que fue de la herramienta FWA que solo fue comparada entre dos autores.

Ilustración 8

Promedio AUC (%) de las herramientas de detección de DeepFakes.

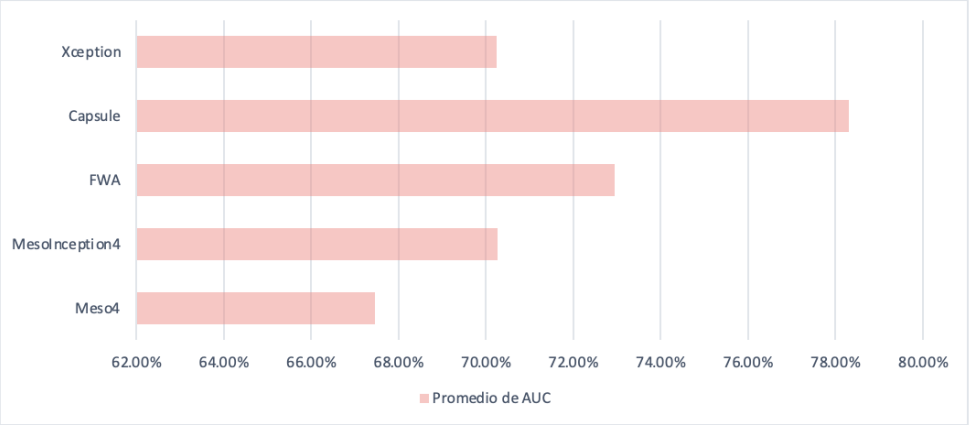


Tabla 21

Promedio de AUC (%) en cada conjunto de datos.

Datasets	*Total*
DFDC	70.07%
FaceForensics++	**78.88%**
Celeb-DF	70.67%

Ilustración 9

Promedio de AUC (%) en cada conjunto de datos.

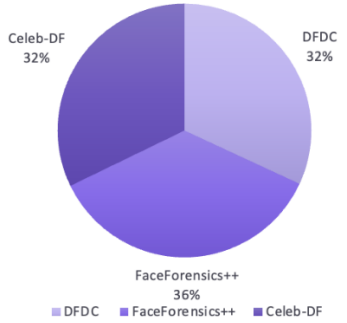


Ilustración 10

Rendimiento de cada herramienta según su autor en el conjunto de datos DFDC

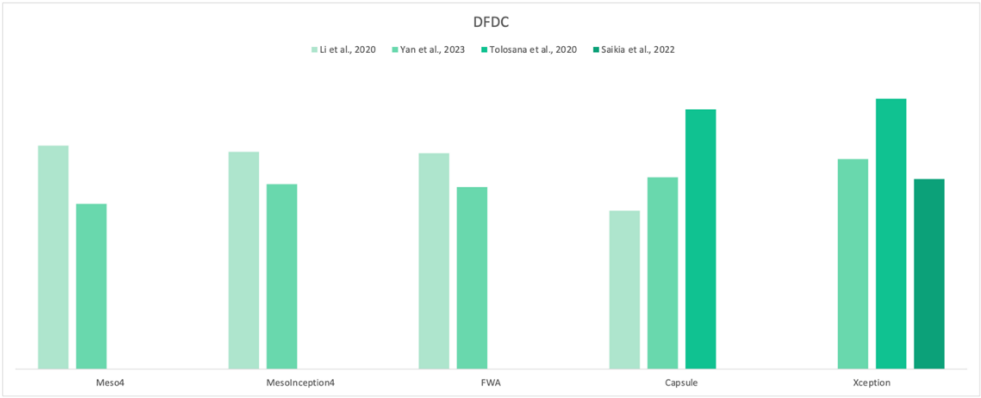


Ilustración 11

Rendimiento de cada herramienta según su autor en el conjunto de datos FaceForensics++

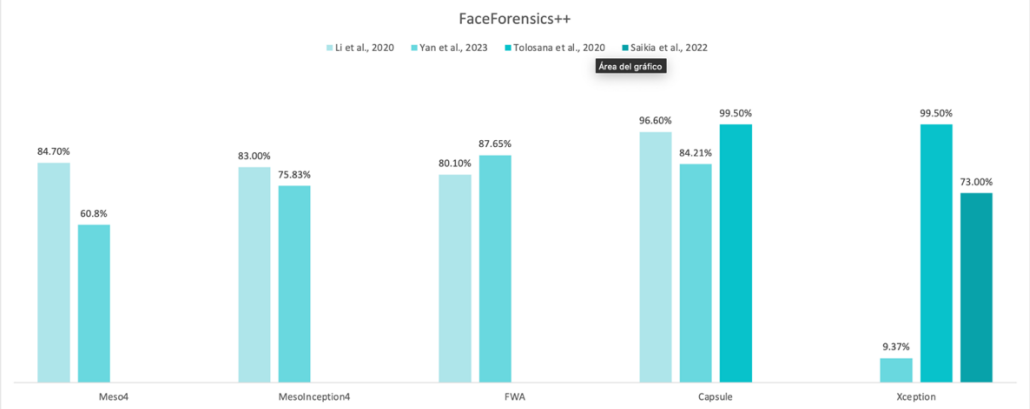
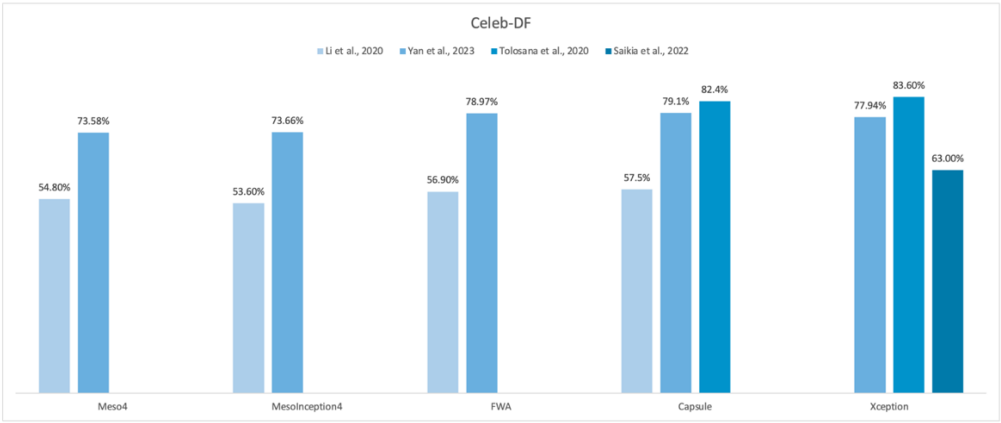


Ilustración 12

Rendimiento de cada herramienta según su autor en el conjunto de datos Celeb-DF



Conclusiones

En conclusión, este estudio ha alcanzado con éxito los objetivos específicos propuestos en relación con la identificación y caracterización del funcionamiento de los métodos empleados para crear DeepFakes, así como la comparación de técnicas y algoritmos utilizados para la detección de estas manipulaciones.

En primer lugar, se logró una comprensión profunda de los métodos utilizados para crear DeepFakes mediante la aplicación de técnicas de entrenamiento de algoritmos de aprendizaje profundo. Esto permitió una evaluación detallada de las etapas clave involucradas en la generación de DeepFakes, incluida la síntesis de imágenes y el procesamiento de audio, lo que contribuyó a una comprensión más completa de su funcionamiento.

En segundo lugar, se llevó a cabo una comparación exhaustiva de las técnicas y algoritmos utilizados para la detección de DeepFakes. Se identificaron diversos enfoques, como la detección de artefactos de compresión, el análisis de movimiento, el análisis de audio y la detección de anomalías en la distribución de píxeles. Esta evaluación permitió identificar las fortalezas y debilidades de cada enfoque, lo que puede ser crucial para el desarrollo de estrategias efectivas de detección en el futuro.

Finalmente, se demostró la funcionalidad de la herramienta tecnológica identificada y analizada sobre contenido multimedia generado por inteligencia artificial. Esto se logró mediante el análisis de los estudios seleccionados que demuestran la aplicación práctica de la herramienta en diversos escenarios y la evaluación de su eficacia en la detección de DeepFakes, lo que proporcionó una validación empírica de su utilidad.

En conjunto, este estudio ha contribuido significativamente a la comprensión y mitigación de los riesgos asociados con DeepFakes, al proporcionar una visión detallada de su creación, detección y las herramientas para abordar este desafío en el mundo digital actual.

El estudio también identificó áreas de mejora y posibles desafíos en la detección de DeepFakes, como la necesidad de considerar diferentes escenarios de generación de videos falsificados y la adaptabilidad de las herramientas a futuras técnicas de manipulación.

Los hallazgos de este experimento contribuyen al cuerpo de conocimiento existente sobre la detección de DeepFakes y ofrecen una base sólida para futuras investigaciones en el campo. A medida que la tecnología de generación de DeepFakes continúa avanzando, la mejora y evolución de estas herramientas de detección se vuelve fundamental para preservar la integridad y veracidad de los contenidos digitales.

En última instancia, se espera que los resultados de este estudio proporcionen una orientación valiosa para los expertos en seguridad digital, investigadores y desarrolladores interesados en la detección y mitigación de DeepFakes, contribuyendo así a la creación de un entorno más confiable y seguro en el ámbito digital.

Recomendaciones

El presente trabajo de titulación ha proporcionado valiosa información sobre la detección de DeepFakes y ha identificado áreas de mejora y posibles desafíos que deben abordarse para fortalecer la efectividad de las herramientas existentes. Basándose en los hallazgos obtenidos, se ofrecen las siguientes recomendaciones:

- Dado el continuo desarrollo y evolución de las técnicas de generación de DeepFakes, se recomienda a investigadores estar al día con las últimas tendencias en la generación de contenido falso permitirá adaptar las herramientas de detección de manera más efectiva.
- Los resultados de este trabajo de titulación han destacado la importancia de contar con conjunto de datos diversos para evaluar el desempeño de las herramientas de detección. Se recomienda expandir y mejorar la variedad de escenarios y contextos de generación de DeepFakes en los conjuntos de datos utilizados. Esto asegurará que las herramientas de detección sean más robustas y versátiles en la detección de falsificaciones.
- Debido a la complejidad y velocidad con la que evolucionan las tecnologías de generación y detección de DeepFakes, se destaca la importancia de una colaboración estrecha entre la industria y la academia. Compartir conocimientos y recursos promoverá el desarrollo conjunto de soluciones más eficientes y efectivas frente a este desafío en constante cambio.
- Es esencial llevar a cabo evaluaciones constantes y rigurosas del desempeño de las herramientas de detección de deepfakes para garantizar su eficacia a lo largo del tiempo, entrenando las herramientas con nuevos conjuntos de datos y escenarios para medir y mejorar el rendimiento de las herramientas existentes.
- Además de desarrollar tecnologías de detección avanzadas, se recomienda fomentar la educación y concienciación sobre el fenómeno de los DeepFakes.

Promover la alfabetización digital y aumentar la comprensión pública sobre los riesgos asociados con la manipulación de contenidos puede ayudar a reducir su impacto y difusión.

En última instancia, se espera que las recomendaciones anteriores guíen a los expertos en seguridad digital, investigadores y desarrolladores interesados en la detección y mitigación de deepfakes. Mediante una colaboración continua y el enfoque en la innovación, se puede contribuir significativamente a la creación de un entorno digital más confiable y seguro para todos los usuarios.

Referencias

Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A Compact Facial Video Forgery Detection Network. 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 1–7. https://doi.org/10.1109/WIFS.2018.8630761

Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S.-N. (2020). Detecting Deep-Fake Videos from Appearance and Behavior. 2020 IEEE International Workshop on Information Forensics and Security (WIFS), 1–6. https://doi.org/10.1109/WIFS49906.2020.9360904

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019). Protecting World Leaders Against Deep Fakes.

Ahmed, S. (2021). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. Telematics and Informatics, 57, 101508. https://doi.org/10.1016/j.tele.2020.101508

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). THE STATE OF DEEPFAKES LANDSCAPE, THREATS, AND IMPACT.

Albahar, M., & Almalki, J. (2019). DEEPFAKES: THREATS AND COUNTERMEASURES SYSTEMATIC REVIEW. https://www.semanticscholar.org/paper/DEEPFAKES%3A-THREATS-AND-COUNTERMEASURES-SYSTEMATIC-Albahar-Almalki/cd1cbbe9b7e5cb47c9f3aaf1b475d4694d9b2492

ASALE, R.-, & RAE. (n.d.-a). Caracterizar | Diccionario de la lengua española. «Diccionario de la lengua española» - Edición del Tricentenario. Retrieved June 17, 2023, from https://dle.rae.es/caracterizar

ASALE, R.-, & RAE. (n.d.-b). Identificar | Diccionario de la lengua española. «Diccionario de la lengua española» - Edición del Tricentenario. Retrieved June 17, 2023, from https://dle.rae.es/identificar

Bateman, J. (2020). Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios.

Brown, R. (2020, February 4). What is Deepfake: How it Works and Spot with Detection Services? Cogitotech. https://www.cogitotech.com/blog/what-is-deepfake-works-detection-services/

Caporusso, N. (2021). Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology (pp. 235–241). https://doi.org/10.1007/978-3-030-51328-3_33

Coccomini, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022). Combining EfficientNet and Vision Transformers for Video Deepfake Detection. In S. Sclaroff, C. Distante, M. Leo, G. M. Farinella, & F. Tombari (Eds.), Image Analysis and Processing – ICIAP 2022 (pp. 219–229). Springer International Publishing. https://doi.org/10.1007/978-3-031-06433-3_19

DeepfakeVFX.com. (2021, November 8). DeepFaceLab—DeepfakeVFX.com. https://www.deepfakevfx.com/downloads/deepfacelab/

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The DeepFake Detection Challenge (DFDC) Dataset (arXiv:2006.07397). arXiv. http://arxiv.org/abs/2006.07397

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). The Deepfake Detection Challenge (DFDC) Preview Dataset (arXiv:1910.08854). arXiv. http://arxiv.org/abs/1910.08854

Gamage, D., Chen, J., & Sasahara, K. (2021). The Emergence of Deepfakes and its Societal Implications: A Systematic Review.

Gavrovska, A. (2022). From puppet-master creation to false detection.

Gosse, C., & Burkell, J. (2020). Politics and porn: How news media characterizes problems presented by deepfakes. Critical Studies in Media Communication, 37(5), 497–511. https://doi.org/10.1080/15295036.2020.1832697

Higgins JPT, T. J., Chandler J, P. M., & Welch VA. (2023). Cochrane Handbook for Systematic Reviews of Interventions version 6.4. https://training.cochrane.org/handbook/current/chapter-01

Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., & Ye, D. (2023). Implicit Identity Driven Deepfake Face Swapping Detection. 4490–4499. https://openaccess.thecvf.com/content/CVPR2023/html/Huang_Implicit_Identity_Driven_Deepfake_Face_Swapping_Detection_CVPR_2023_paper.html

Hussain, S., Neekhara, P., Jere, M., Koushanfar, F., & McAuley, J. (2021). Adversarial Deepfakes: Evaluating Vulnerability of Deepfake Detectors to Adversarial Examples.

2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 3347–3356. https://doi.org/10.1109/WACV48630.2021.00339

Intel. (2022, November 14). Intel Introduces Real-Time Deepfake Detector. Intel. https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html

Jiang, L., Wu, W., Qian, C., & Loy, C. C. (2022). DeepFakes detection: The DeeperForensics dataset and challenge. Handbook of Digital Face Manipulation and Detection, 303.

Kerner, C., & Risse, M. (2021). Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. Moral Philosophy and Politics, 8(1), 81–108. https://doi.org/10.1515/mopp-2020-0024

Khalid, H., & Woo, S. S. (2020). OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2794–2803. https://doi.org/10.1109/CVPRW50498.2020.00336

Kietzmann, J., Lee, L., McCarthy, I., & Kietzmann, T. (2019). Deepfakes: Trick or treat? Business Horizons, 63. https://doi.org/10.1016/j.bushor.2019.11.006

Kwok, A. O. J., & Koh, S. G. M. (2021). Deepfake: A social construction of technology perspective. Current Issues in Tourism, 24(13), 1798–1802. https://doi.org/10.1080/13683500.2020.1738357

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 3204–3213. https://doi.org/10.1109/CVPR42600.2020.00327

Lin, C., Deng, J., Hu, P., Shen, C., Wang, Q., & Li, Q. (2022). Towards Benchmarking and Evaluating Deepfake Detection (arXiv:2203.02115). arXiv. http://arxiv.org/abs/2203.02115

Mahmud, B. U., & Sharmin, A. (2023). Deep Insights of Deepfake Technology: A Review (arXiv:2105.00192). arXiv. http://arxiv.org/abs/2105.00192

Masood, M., Nawaz, M., Malik, K. M., Javed, A., & Irtaza, A. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward (arXiv:2103.00484). arXiv. http://arxiv.org/abs/2103.00484

Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. ACM Computing Surveys, 54(1), 7:1-7:41. https://doi.org/10.1145/3425780

Montserrat, D. M., Hao, H., Yarlagadda, S. K., Baireddy, S., Shao, R., Horvath, J., Bartusiak, E., Yang, J., Guera, D., Zhu, F., & Delp, E. J. (2020). Deepfakes Detection with Automatic Face Weighting. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2851–2859. https://doi.org/10.1109/CVPRW50498.2020.00342

Müller, N. M., Dieckmann, F., Czempin, P., Canals, R., Böttinger, K., & Williams, J. (2021). Speech is Silver, Silence is Golden: What do ASVspoof-trained Models Really Learn? (arXiv:2106.12914). arXiv. http://arxiv.org/abs/2106.12914

Newman, L. H. (2019, February 11). A New Tool Protects Videos From Deepfakes and Tampering. Wired. https://www.wired.com/story/amber-authenticate-video-validation-blockchain-tampering-deepfakes/

Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Use of a Capsule Network to Detect Fake Images and Videos (arXiv:1910.12467). arXiv. http://arxiv.org/abs/1910.12467

Nguyen, T., Nguyen, C. M., Nguyen, T., Duc, T., & Nahavandi, S. (2019). Deep Learning for Deepfakes Creation and Detection: A Survey.

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

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., Jiang, J., Zhang, S., Wu, P., Zhou, B., & Zhang, W. (2021). DeepFaceLab: Integrated, flexible and extensible face-swapping framework (arXiv:2005.05535). arXiv. https://doi.org/10.48550/arXiv.2005.05535

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2018). FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces (arXiv:1803.09179; Version 1). arXiv. http://arxiv.org/abs/1803.09179

Saikia, P., Dholaria, D., Yadav, P., Patel, V., & Roy, M. (2022). A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features. 2022 International Joint Conference on Neural Networks (IJCNN), 1–7. https://doi.org/10.1109/IJCNN55064.2022.9892905

Sensity AI. (2023, January 29). Top Deepfake Detection Solution | New AI Image Detection. https://sensity.ai/deepfake-detection/

Silva, S. H., Bethany, M., Votto, A. M., Scarff, I. H., Beebe, N., & Najafirad, P. (2022). Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. Forensic Science International: Synergy, 4, 100217. https://doi.org/10.1016/j.fsisyn.2022.100217

The World Economic Forum. (2023). Sensity.ai. World Economic Forum. https://www.weforum.org/organizations/sensity-ai/

Tolosana, R., Romero-Tapiador, S., Fierrez, J., & Vera-Rodriguez, R. (2020). DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance (arXiv:2004.07532). arXiv. http://arxiv.org/abs/2004.07532

Tran, V.-N., Lee, S.-H., Le, H.-S., & Kwon, K.-R. (2021). High Performance DeepFake Video Detection on CNN-Based with Attention Target-Specific Regions and Manual Distillation Extraction. Applied Sciences, 11, 7678. https://doi.org/10.3390/app11167678

Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101

Walczyna, T., & Piotrowski, Z. (2023). Quick Overview of Face Swap Deep Fakes. Applied Sciences, 13(11), Article 11. https://doi.org/10.3390/app13116711

Weerawardana, M., & Fernando, T. (2021). Deepfakes Detection Methods: A Literature Survey. 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS), 76–81. https://doi.org/10.1109/ICIAfS52090.2021.9606067

Yadav, D., & Salmani, S. (2019). Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network. 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 852–857. https://doi.org/10.1109/ICCS45141.2019.9065881

Yan, Z., Zhang, Y., Yuan, X., Lyu, S., & Wu, B. (2023). DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection (arXiv:2307.01426). arXiv. http://arxiv.org/abs/2307.01426

Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., & Loy, C. C. (2022). CelebV-HQ: A Large-Scale Video Facial Attributes Dataset. In S. Avidan, G. Brostow, M. Cissé,

G. M. Farinella, & T. Hassner (Eds.), Computer Vision – ECCV 2022 (pp. 650–667). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-20071-7_38

DECLARACIÓN Y AUTORIZACIÓN

Yo, **Costa Mora, Brittany Valeria** con C.C: # **0950924217** autora del trabajo de titulación: **Estudio metodológico de las herramientas tecnológicas actuales para detectar y reconocer deepfakes, abordando la creciente amenaza de la manipulación de contenido multimedia generado por inteligencia artificial,** previo a la obtención del título de **Ingeniero en Ciencias de la Computación** en la Universidad Católica de Santiago de Guayaquil.

1.- Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar a la SENESCYT en formato digital una copia del referido trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.

2.- Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de propiedad intelectual vigentes.

Guayaquil, 8 de septiembre de 2023

Brittany Costa

_____

Nombre: **Costa Mora, Brittany Valeria**
C.C: **0950924217**

REPOSITORIO NACIONAL EN CIENCIA Y TECNOLOGÍA

FICHA DE REGISTRO DE TESIS/TRABAJO DE TITULACIÓN

TEMA Y SUBTEMA:	Estudio metodológico de las herramientas tecnológicas actuales para detectar y reconocer deepfakes, abordando la creciente amenaza de la manipulación de contenido multimedia generado por inteligencia artificial.		
AUTOR(ES)	Costa Mora, Brittany Valeria		
REVISOR(ES)/TUTOR(ES)	Ing. Castro Aguilar, Gilberto Castro		
INSTITUCIÓN:	Universidad Católica de Santiago de Guayaquil		
FACULTAD:	Ingeniería		
CARRERA:	Ingeniería en Ciencias de la Computación		
TITULO OBTENIDO:	Ingeniero en Ciencias de la Computación		
FECHA DE PUBLICACIÓN:	8 de septiembre de 2023	**No. DE PÁGINAS:**	60
ÁREAS TEMÁTICAS:	Inteligencia Artificial, Aprendizaje Profundo, DeepFakes		
PALABRAS CLAVES/ KEYWORDS:	DeepFakes, inteligencia artificial, detección, AUC, aprendizaje automático, aprendizaje profundo.		

RESUMEN/ABSTRACT:

Este trabajo de titulación busca abordar la creciente amenaza de los DeepFakes en la era digital, donde la manipulación de contenido multimedia generado por inteligencia artificial puede dificultar la distinción entre información real y falsa. El objetivo general de esta investigación ha sido identificar, analizar y evaluar las herramientas tecnológicas utilizadas para detectar y reconocer DeepFakes, con el propósito de enfrentar este desafío de manera efectiva.

Para lograrlo, se ha diseñado una ruta de trabajo que proporciona claridad en los procesos y pasos a seguir, permitiendo así la elaboración de una metodología sólida y bien fundamentada. La elección de la métrica AUC como criterio de evaluación ha permitido medir de manera global la capacidad discriminativa de las herramientas, asegurando una comparación objetiva y justa.

Los resultados obtenidos de la comparación y evaluación de cinco herramientas de detección de DeepFakes han revelado que Meso4 y Capsule se destacaron con los puntajes más altos en diferentes estudios, sobresaliendo especialmente en el conjunto de datos FaceForensics++. Además, se han identificado áreas de mejora y desafíos en la detección de DeepFakes, resaltando la importancia de considerar diferentes escenarios de generación de videos falsificados y la adaptabilidad de las herramientas a futuras técnicas de manipulación.

Este estudio aporta una valiosa contribución al campo de la detección de DeepFakes, ofreciendo una base sólida para futuras investigaciones en esta área. Los resultados y conclusiones obtenidos serán de gran interés para personas interesadas en la detección y mitigación de DeepFakes.

La implementación de estas herramientas de detección y la mejora continua de las mismas permitirá preservar la integridad y veracidad de los contenidos digitales en un contexto de rápida evolución tecnológica. De esta manera, se espera hacer frente a la amenaza de los DeepFakes y promover la confianza en la información en la sociedad actual.

ADJUNTO PDF:	☒ SI	☐ NO
CONTACTO CON AUTOR/ES:	**Teléfono:**+593-967-890681	E-mail: **brittanycosta522@gmail.com**
CONTACTO CON LA INSTITUCIÓN (C00RDINADOR DEL PROCESO UTE)::	**Toala Quimí, Edison José**	
	Teléfono: +593-990-976776	
	E-mail: edison.toala@cu.ucsg.edu.ec	

SECCIÓN PARA USO DE BIBLIOTECA

Nº. DE REGISTRO (en base a datos):	
Nº. DE CLASIFICACIÓN:	
DIRECCIÓN URL (tesis en la web):	