



Escuela
Politécnica
Superior

Prevención de Suplantación Facial mediante GANs y Detección de Profundidad en Reconocimiento Biométrico

Máster Universitario en
**Ciber
Seg** Ciberseguridad

Trabajo Fin de Máster

Autor:

José Burgos Miras

Tutor/es:

Ester Martínez Martín

Mayo 2025

Prevención de Suplantación Facial mediante GANs y Detección de Profundidad en Reconocimiento Biométrico

Subtítulo del proyecto

Autor

José Burgos Miras

Tutor/es

Ester Martínez Martín

Departamento de Ciencia de la Computación e Inteligencia Artificial



Máster Universitario en Ciberseguridad



Escuela
Politécnica
Superior



Universitat d'Alacant
Universidad de Alicante

ALICANTE, Mayo 2025

Preámbulo

Este Trabajo Fin de Máster representa la culminación de mi formación en el Máster Universitario en Ciberseguridad de la Universidad de Alicante, así como un paso más en mi desarrollo personal y académico en un área que me apasiona profundamente. Tras haber explorado previamente las aplicaciones del aprendizaje profundo en el ámbito de la visión por ordenador, este proyecto me ha permitido orientar mi interés hacia un problema crítico en el campo de la ciberseguridad biométrica: la detección de ataques por suplantación facial.

La elección del tema ha estado motivada por el deseo de profundizar en soluciones prácticas ante amenazas emergentes, aprovechando técnicas de inteligencia artificial tanto generativas como analíticas. El enfoque adoptado ha combinado el uso de redes generativas antagónicas (GANs) y el análisis de mapas de profundidad, con el objetivo de evaluar su potencial como herramientas complementarias en la verificación facial.

Quiero expresar mi más sincero agradecimiento a la Dra. Ester Martínez Martín, del Departamento de Ciencia de la Computación e Inteligencia Artificial, por su continua implicación, confianza y acompañamiento a lo largo de este y otros proyectos formativos. Su visión crítica, su apoyo técnico y su cercanía han sido determinantes en la definición y ejecución de este trabajo.

Del mismo modo, agradezco al profesorado del máster por el conocimiento compartido durante el curso, y a mi entorno personal por el respaldo incondicional que ha facilitado mi dedicación durante estos meses. Este proyecto no solo cierra una etapa académica, sino que refuerza mi compromiso con el aprendizaje continuo en el ámbito de la ciberseguridad y la inteligencia artificial.

Agradecimientos

En primer lugar, deseo expresar mi más sincero agradecimiento a mi tutora, la Dra. Ester Martínez Martín, por su constante apoyo, implicación y orientación experta durante el desarrollo de este trabajo. Su experiencia, su visión crítica y su cercanía han sido fundamentales no solo para la ejecución técnica del proyecto, sino también para mi crecimiento como investigador.

También agradezco a la Universidad de Alicante por brindarme el entorno académico y los recursos necesarios para llevar a cabo esta investigación, así como al Departamento de Ciencia de la Computación e Inteligencia Artificial por facilitarme el acceso a su infraestructura computacional y herramientas colaborativas, que han sido clave para la parte experimental del estudio.

No puedo dejar de agradecer a mi familia por su apoyo incondicional, su paciencia y su confianza constante en mí. En los momentos de mayor exigencia y dificultad, su presencia ha sido mi mayor fuente de energía y motivación.

Finalmente, extiendo mi gratitud a todos los compañeros, docentes y profesionales del Máster Universitario en Ciberseguridad por los conocimientos compartidos, las discusiones enriquecedoras y el estímulo continuo a lo largo de este proceso formativo. Este trabajo es también el reflejo de ese recorrido conjunto.

*A mis padres, por su amor constante y su ejemplo incansable.
A mi hermana, que ha sido una fuente inesperada de apoyo y orgullo.
Ojalá este trabajo esté a la altura de lo que esperas de tu hermano mayor.*

Una sola vulnerabilidad es todo lo que un atacante necesita.

Window Snyder.

Índice general

1	Introducción	1
2	Estado del Arte	5
2.1	Introducción	5
2.2	Ciberseguridad e Inteligencia Artificial	5
2.2.1	Aplicaciones Generales de Inteligencia Artificial (IA) en Ciberseguridad	5
2.2.2	Marcos de Referencia en Ciberseguridad	6
2.2.3	IA y Computación Cuántica en Ciberseguridad	6
2.2.4	IA en Criptografía	6
2.2.4.1	Criptografía de Texto con Aprendizaje Profundo	7
2.2.4.2	Aplicaciones de Redes Neuronales en Criptografía	7
2.2.4.3	Aplicaciones Multidisciplinarias de IA en Criptografía	8
2.2.4.4	Sistemas Avanzados de Cifrado con Redes Neuronales y Sistemas Caóticos	9
2.3	Redes Adversariales Generativas (GAN) en Ciberseguridad	10
2.3.1	Aplicaciones de las Redes Adversariales Generativas (Generative Adversarial Networks) (GAN)s	10
2.3.2	Explicabilidad y Transparencia en las GANs	10
2.4	Limitaciones y Oportunidades Futuras	11
2.5	Conclusiones del Estado del Arte	11
3	Fundamentos Teóricos	13
3.1	Introducción	13
3.2	Conceptos Fundamentales en Ciberseguridad	13
3.2.1	Definición de Ciberseguridad	13
3.2.1.1	Importancia de la Ciberseguridad	14
3.2.2	Principales Amenazas y Vulnerabilidades	14
3.2.2.1	Malware	14
3.2.2.2	Ransomware	14
3.2.2.3	Phishing	14
3.2.2.4	Ataques Adversariales	15
3.2.2.5	Impacto en Sistemas de Autenticación Biométrica	15
3.2.3	Modelos de Seguridad en Sistemas de Información	15
3.2.3.1	Tríada CIA: Confidencialidad, Integridad y Disponibilidad	15
3.2.3.2	Modelo de Confianza Cero (Zero Trust)	16
3.2.3.3	Defensa en Profundidad	16
3.2.4	Marcos Regulatorios y Normativas de Seguridad	16
3.2.4.1	Reglamento General de Protección de Datos (Reglamento General de Protección de Datos (GDPR))	17

3.2.4.2	ISO/IEC 27001	17
3.2.4.3	Marco de Ciberseguridad del Instituto Nacional de Estándares y Tecnología (National Institute of Standards and Technology) (NIST)	17
3.2.4.4	Impacto en la Implementación de IA en Ciberseguridad	17
3.3	Inteligencia Artificial y Ciberseguridad	18
3.3.1	Conceptos Básicos de IA y Aprendizaje Automático	18
3.3.2	Uso de IA en Ciberseguridad	18
3.3.3	Ataques Adversariales en IA	19
3.4	Redes Adversariales Generativas (GANs)	20
3.4.1	Concepto y Arquitectura de las GANs	20
3.4.2	Aplicaciones de GANs en Ciberseguridad	21
3.4.2.1	Usos ofensivos: ataques mediante GANs	21
3.4.2.2	Usos defensivos: protección mediante GANs	22
3.4.3	Retos y Desafíos de las GANs en Seguridad	22
3.5	Síntesis y Relación con la Solución Propuesta	23
4	Metodología	25
4.1	Enfoque Metodológico	25
4.2	Desarrollo de un Sistema Basado en GANs	25
4.3	Desarrollo de un Sistema Basado en GANs	26
4.3.1	Diseño del Generador y Discriminador	26
4.3.2	Preparación del Dataset	28
4.3.2.1	Preparación del conjunto de datos	28
4.3.2.1.1	Conjunto de datos de validación experimental	28
4.3.2.1.2	Conjunto de datos para entrenamiento del generador	28
4.3.2.1.3	Conjunto de datos para clasificación de emociones	30
4.3.3	Entrenamiento del Modelo GAN	32
4.3.3.1	Etapa 1: Validación funcional con Instituto Nacional de Estándares y Tecnología Modificado (Modified National Institute of Standards and Technology) (MNIST)	36
4.3.3.2	Etapa 2: Generación de rostros desde vídeos de ataques por impresión	36
4.3.3.3	Etapa 3: Generación sobre expresiones faciales (Conjunto de Datos de Reconocimiento de Emociones Faciales (Facial Expression Recognition 2013) (FER2013))	37
4.3.3.4	Etapa 4: Evaluación sistemática y control de sobreajuste	37
4.3.4	Estrategia de Evaluación	37
4.4	Sistema de Detección Basado en Profundidad	38
4.4.1	Procesamiento de Imágenes con MiDaS	39
4.4.2	Extracción de Región de Interés (Region of Interest) (ROI) y Análisis de Textura	41
4.4.3	Comparación entre Imagen Real y Fotocopia	43
4.5	Herramientas y Entorno de Desarrollo	47
4.5.1	Librerías y Frameworks	47

4.5.2	Entorno de Ejecución	47
4.5.3	Estructura del Proyecto	47
4.5.4	Intercambio de Archivos	49
4.5.5	Reproducibilidad	49
4.5.6	Visualización y Guardado de Resultados	49
4.5.7	Automatización y Scripts Personalizados	49
4.6	Limitaciones y Desafíos Encontrados	50
4.6.1	Dificultades en la Línea GAN	50
4.6.2	Limitaciones del Análisis con MiDaS	50
4.6.3	Restricciones Técnicas y del Entorno de Ejecución	51
5	Resultados	53
5.1	Evaluación del Sistema GAN	53
5.1.1	Etapa 2: Análisis por resolución sobre ataques por impresión (gan2.py)	54
5.1.1.1	Evaluación cuantitativa: precisión del discriminador.	54
5.1.1.2	Evaluación cualitativa: muestras generadas.	57
5.1.2	Etapa 3: Evaluación con datos de emociones (gan3.py)	58
5.1.3	Etapa 4: Evaluación cruzada con conjunto de test (gan4.py)	61
5.1.4	Consideraciones globales sobre el aprendizaje adversarial	63
5.2	Evaluación del Análisis de Profundidad con MiDaS	64
5.2.1	Mapas de Profundidad y Visualización Directa	65
5.2.2	Histogramas de Profundidad y Análisis de Textura	68
5.2.3	Histogramas de la Imagen Completa	68
5.2.3.1	Histogramas de la Región de Interés (ROI)	70
5.2.3.2	Conclusiones del análisis de histogramas	71
5.2.4	Comparación entre Métodos de Extracción de ROI	71
5.2.4.1	Método 1: Detección por contornos	72
5.2.4.2	Método 2: Detección mediante <i>Haar cascades</i>	72
5.2.4.3	Comparación Visual: Pepe1	73
5.2.4.4	Comparación Visual: Pepe2	73
5.2.5	Indicadores Cuantitativos y Observaciones Generales	74
5.2.5.1	Métricas empleadas	75
5.2.5.2	Visualización comparativa	75
5.2.5.3	Resumen numérico de resultados	77
5.2.5.4	Ánalisis de los resultados	77
5.2.5.5	Conclusiones y líneas futuras	78
5.3	Reflexión Final sobre los Resultados	78
6	Discusión	81
6.1	Interpretación General de los Resultados	81
6.1.1	Análisis de Resultados en el Contexto del Proyecto	81
6.1.2	Relación con Problemas de Verificación Facial	81
6.2	Limitaciones y Desafíos Encontrados	82
6.2.1	Dificultades en el Entrenamiento GAN	82
6.2.2	Restricciones del Análisis de Profundidad	82

6.3	Robustez, Fiabilidad y Generalización	83
6.3.1	Robustez del Sistema GAN	83
6.3.2	Capacidad de Generalización del Modelo de Profundidad	84
6.4	Implicaciones Metodológicas y Técnicas	84
6.4.1	Lecciones Aprendidas sobre Aprendizaje Adversarial	84
6.4.2	Valor del Análisis Estructural Basado en Profundidad	85
6.5	Síntesis de Aportes	86
6.5.1	Valor Científico y Técnico del Trabajo	86
6.5.2	Relevancia dentro del Campo de la Verificación Facial	86
7	Conclusiones	89
7.1	Cumplimiento de Objetivos	89
7.2	Principales Contribuciones	90
7.3	Impacto de los Resultados	91
7.4	Limitaciones Detectadas	92
7.5	Cierre y Perspectiva	93
8	Trabajos Futuros	95
8.1	Ampliación y Validación Experimental	95
8.2	Mejoras del Sistema GAN	95
8.2.1	Ajuste de Arquitectura y Hiperparámetros	95
8.2.2	Uso del Generador como Fuente de Datos	96
8.3	Fortalecimiento del Análisis de Profundidad	96
8.3.1	Evaluación con Datos Más Diversos	96
8.3.2	Automatización y Optimización	96
8.4	Despliegue en Entornos Reales	97
8.4.1	Funcionamiento en Tiempo Real	97
8.4.2	Desarrollo de Herramienta Interactiva	97
8.5	Consideraciones Éticas y de Seguridad	97
Bibliografía		99
Lista de Acrónimos y Abreviaturas		105

Índice de figuras

4.1	Arquitectura del generador basada en capas densas, normalización por lotes y activaciones LeakyReLU.	27
4.2	Arquitectura del discriminador basada en capas densas, activaciones LeakyReLU y salida sigmoid.	27
4.3	Ejemplos de imágenes del dataset MNIST utilizadas para validación experimental.	28
4.4	Estructura real del directorio de trabajo para la versión de la GAN entrenada con el conjunto de datos de ataques por impresión obtenido de Kaggle.	29
4.5	Secuencia de frames extraídos del vídeo Print01.	29
4.6	Secuencia de frames extraídos del vídeo Print10.	30
4.7	Secuencia de frames extraídos del vídeo Print20.	30
4.8	Ejemplos de imágenes por clase en el dataset de emociones FER2013.	31
4.9	Distribución de imágenes por clase en el conjunto de entrenamiento.	31
4.10	Distribución de imágenes por clase en el conjunto de prueba.	32
4.11	Comparativa entre las distribuciones del conjunto de entrenamiento y prueba.	32
4.12	Evolución de las imágenes generadas por la GAN a lo largo del entrenamiento, desde la época 0 hasta la 900. Puede observarse una mejora progresiva en la definición y estructura de los rostros sintéticos.	33
4.13	Comparativa de la calidad de las imágenes generadas por la GAN en tres momentos clave del entrenamiento (épocas 0, 400 y 900). La mejora en la coherencia visual y los rasgos faciales es evidente.	34
4.14	Imagen generada por la GAN en la época 0. Aún no se distinguen rasgos faciales definidos.	35
4.15	Imagen generada por la GAN en la época 400. Comienzan a apreciarse estructuras faciales.	35
4.16	Imagen generada por la GAN en la época 900. Los rostros sintéticos presentan mayor definición y coherencia.	36
4.17	Flujo simplificado del procesamiento de una imagen con MiDaS.	40
4.18	Imagen de entrada (izquierda) y su respectivo mapa de profundidad coloreado (derecha) generado con MiDaS.	40
4.19	Ejemplo de extracción de ROI: imagen original (izquierda), máscara facial generada (centro) y ROI aplicada sobre el mapa de profundidad (derecha).	42
4.20	Histogramas de profundidad de las regiones de interés. Izquierda: imagen real. Derecha: fotocopia.	43
4.21	Histogramas de profundidad en la región facial. Izquierda: imagen original. Derecha: fotocopia.	45
4.22	Imagen original (izquierda) y su versión fotocopiada (derecha) empleadas en el análisis de profundidad.	46

4.23 Comparativa de ROIs: recortes de profundidad (izquierda) y máscaras faciales Haar (derecha), para imagen real y fotocopia.	46
5.1 Precisión del discriminador - Resolución 28x28.	55
5.2 Precisión del discriminador - Resolución 52x52.	56
5.3 Comparativa de precisión entre tres ejecuciones independientes con resolución 96x96.	56
5.4 Comparativa de precisión para las tres resoluciones evaluadas.	57
5.5 Imágenes generadas por la GAN en distintas épocas (resolución 96x96).	57
5.6 Evolución de la precisión del discriminador para las tres ejecuciones del script <code>gan3.py</code> utilizando el dataset FER2013 (48x48).	58
5.7 Evolución de las imágenes generadas por <code>gan3.py</code> desde la época 0 hasta la 900. Puede apreciarse una mejora notable en la estructura facial y la coherencia visual.	59
5.8 Comparativa visual de las imágenes generadas por el modelo entrenado con <code>gan3.py</code> en las épocas 0, 400 y 900. Se observa un incremento progresivo en la calidad de los rostros sintetizados.	59
5.9 Imágenes generadas en la época 0 con <code>gan3.py</code> . Se observa ruido aleatorio sin rasgos definidos.	60
5.10 Imágenes generadas en la época 400 con <code>gan3.py</code> . Comienzan a emerge estructuras faciales básicas.	60
5.11 Imágenes generadas en la época 900 con <code>gan3.py</code> . Se distinguen claramente rostros con expresiones faciales.	61
5.12 Precisión del discriminador sobre el conjunto de test en las tres ejecuciones de <code>gan4.py</code>	62
5.13 Distribución de las clases emocionales en el conjunto de entrenamiento (FER2013).	62
5.14 Comparativa visual entre muestras del conjunto de entrenamiento y del conjunto de test.	63
5.15 Comparativa entre imagen real y fotocopia en el caso ester5	65
5.16 Comparativa entre imagen real y fotocopia en el caso pepe3	66
5.17 Comparativa entre imagen real y fotocopia en el caso ester2 (exterior).	67
5.18 Comparativa entre imagen real y fotocopia en el caso pepe1	67
5.19 Histogramas de profundidad para la imagen real y la fotocopia en el caso pepe1	69
5.20 Histogramas de profundidad para la imagen real y la fotocopia en el caso ester5	69
5.21 Comparativa de histogramas de profundidad en la región facial en el caso ester2 (exterior).	70
5.22 Comparativa de histogramas de profundidad en la región facial en el caso ester4 (interior).	71
5.23 Imágenes originales y sus correspondientes fotocopias en los casos pepe1 y pepe2	72
5.24 Comparativa de ROIs extraídas en el caso pepe1	73
5.25 Comparativa de ROIs extraídas en el caso pepe2	74
5.26 Comparación de la variación de profundidad entre imágenes reales y fotocopias.	75
5.27 Comparación de la textura (varianza del Laplaciano) entre imágenes reales y fotocopias.	76

5.28 Comparación de la desviación típica de profundidad entre imágenes reales y fotocopias.	76
---	----

Índice de tablas

5.1 Resumen cuantitativo de las métricas calculadas en imágenes reales y fotocopias. 77

Índice de Códigos

1 Introducción

En un mundo cada vez más digitalizado y conectado, la seguridad de los sistemas de identificación se ha convertido en una prioridad crucial. El reconocimiento facial, como una de las formas más avanzadas y extendidas de autenticación biométrica, desempeña un papel fundamental en la protección de información sensible, el control de accesos y la prevención de actividades fraudulentas. Sin embargo, esta misma dependencia de la biometría presenta nuevos desafíos, especialmente frente a la creciente sofisticación de los ataques adversariales, donde imágenes manipuladas o generadas artificialmente buscan comprometer la integridad de los sistemas.

En el ámbito de la ciberseguridad, los ataques adversariales constituyen una amenaza significativa. Estos ataques, diseñados para explotar vulnerabilidades en sistemas basados en aprendizaje automático, no solo buscan engañar a los modelos de reconocimiento, sino también socavar la confianza del usuario en la tecnología. Ante esta problemática, se hace evidente la necesidad de desarrollar soluciones innovadoras y robustas que no solo fortalezcan la defensa contra estos ataques, sino que también aprovechen los avances en Inteligencia Artificial (IA) para mejorar la precisión y la seguridad de los sistemas.

En este contexto, las redes adversariales generativas (GAN, por sus siglas en inglés) han emergido como una tecnología transformadora. Originalmente concebidas para generar imágenes sintéticas realistas, las GAN han evolucionado hacia aplicaciones más sofisticadas, incluyendo la detección de manipulaciones adversariales y la mejora de modelos biométricos. Las GAN se componen de dos redes neuronales principales: un generador, que crea imágenes sintéticas, y un discriminador, que aprende a distinguir entre imágenes reales y generadas. Esta dinámica de competencia entre las redes permite no solo generar datos de alta calidad, sino también desarrollar sistemas de evaluación más robustos y precisos.

Este Trabajo Final de Máster (TFM) aborda la problemática de la seguridad en sistemas de reconocimiento facial mediante el diseño e implementación de un modelo basado en GAN, capaz de enfrentar los retos planteados por los ataques adversariales. Asimismo, se ha integrado una fase de análisis estructural a través del uso de modelos preentrenados como Estimación de Profundidad Monocular (Monocular Depth Estimation) (MiDaS)¹. Birkl y cols. (2023), con el fin de estimar mapas de profundidad que permitan detectar patrones característicos en imágenes manipuladas o impresas. Además, se ha propuesto el diseño de una interfaz gráfica como mecanismo de interacción práctica con el sistema, aunque no ha sido completamente desplegada.

¹MiDaS es un modelo de estimación de profundidad monocular desarrollado por Intel, que permite obtener mapas de profundidad a partir de imágenes Rojo Verde Azul (Red Green Blue) (RGB).

Motivación del Proyecto

La motivación principal de este proyecto radica en la necesidad de reforzar los sistemas de identificación biométrica frente a un panorama de amenazas en constante evolución. La creciente adopción de la biometría facial en aplicaciones como el control de acceso, la banca en línea y los sistemas de vigilancia, exige niveles de seguridad cada vez más altos. Sin embargo, la aparición de tecnologías como las GAN, también ha dado lugar a nuevas formas de ataque, incluyendo la generación de rostros falsos que son indistinguibles para sistemas tradicionales. Por lo tanto, desarrollar herramientas que permitan no solo prevenir estos ataques, sino también fortalecer la confianza en las soluciones biométricas, se ha convertido en una prioridad estratégica.

Además, este proyecto responde a un interés personal por explorar la intersección entre Inteligencia Artificial y ciberseguridad, combinando conocimientos avanzados en aprendizaje profundo con las necesidades prácticas del ámbito de la protección digital. Este TFM busca no solo aportar una solución funcional, sino también avanzar en el entendimiento de cómo las redes adversariales pueden ser utilizadas tanto para defender como para comprometer sistemas críticos.

Objetivos del Proyecto

El objetivo general de este proyecto es desarrollar un sistema innovador de identificación facial basado en redes adversariales generativas, que sea capaz de generar imágenes realistas, detectar manipulaciones adversariales y resistir ataques diseñados para comprometer su integridad. Este objetivo se desglosa en los siguientes objetivos específicos:

- Diseñar e implementar un modelo de aprendizaje profundo que integre componentes generativos y discriminativos, optimizados para trabajar con datos de diferentes características y resoluciones.
- Investigar el impacto de distintas configuraciones y parámetros del modelo en la calidad y precisión de los resultados generados, utilizando métricas clave de evaluación.
- Desarrollar un flujo de preprocesamiento de datos que garantice la calidad y representatividad de los mismos para un entrenamiento eficiente y fiable.
- Entrenar el modelo utilizando estrategias avanzadas que minimicen problemas comunes, como la inestabilidad en el entrenamiento, y maximizando el desempeño del sistema.
- Integrar un análisis de mapas de profundidad empleando modelos preentrenados (como MiDaS) para detectar diferencias estructurales entre imágenes reales y manipuladas.
- Identificar los requisitos para una posible interfaz de usuario que facilite la evaluación práctica del sistema, planteando su desarrollo como trabajo futuro.
- Documentar las limitaciones y desafíos técnicos encontrados, proponiendo posibles mejoras y extensiones futuras del sistema desarrollado.

Al final de este trabajo, se espera contribuir de manera significativa al desarrollo de sistemas de identificación facial seguros, estableciendo una base sólida para futuras investigaciones en el campo de las redes adversariales generativas y su aplicación en la seguridad biométrica.

2 Estado del Arte

2.1 Introducción

La ciberseguridad y la inteligencia artificial (IA) han emergido como áreas fundamentales en la protección de sistemas digitales frente a amenazas cada vez más sofisticadas. Diversos estudios recientes destacan la integración de IA en aplicaciones de ciberseguridad, con enfoques que van desde la protección automatizada hasta el uso de modelos avanzados para la detección de amenazas.

En este contexto, las GANs han demostrado ser herramientas clave, capaces de generar datos sintéticos, detectar manipulaciones adversariales y mejorar modelos de aprendizaje automático. Este capítulo está destinado a explorar los avances más relevantes en la integración de IA y ciberseguridad, con especial énfasis en el uso de GANs.

2.2 Ciberseguridad e Inteligencia Artificial

2.2.1 Aplicaciones Generales de IA en Ciberseguridad

A comprehensive study of artificial intelligence and cybersecurity on bitcoin, cryptocurrency, and banking system Choithani y cols. (2024) analiza cómo la IA ha transformado la protección en sistemas financieros, criptomonedas y bancarios. Este artículo identifica aplicaciones como el uso de Máquina de Vectores de Soporte (Support Vector Machine) (SVM), Redes Neuronales Artificiales (Artificial Neural Networks) (ANN), Memoria a Largo y Corto Plazo (Long Short-Term Memory) (LSTM) y Unidad Recurrente Gated (Gated Recurrent Unit) (GRU) para predecir tendencias de precios y gestionar riesgos, enfrentando desafíos como la volatilidad y la falta de datos a largo plazo. También destaca el uso de blockchain para transacciones seguras y tecnologías de prevención como Sistema de Prevención de Intrusiones (Intrusion Prevention System) (IPS), Sistema de Detección de Intrusiones (Intrusion Detection System) (IDS) y Gestión de Información y Eventos de Seguridad (Security Information and Event Management) (SIEM).

Artificial intelligence for cybersecurity: Current trends and future challenges Meghna Manoj Nair (2024) proporciona una visión integral sobre el uso de la IA en ciberseguridad, enfatizando sus capacidades para detectar y prevenir ataques rápidamente, automatizar el análisis de patrones y proteger infraestructuras críticas en sectores como la sanidad y los servicios públicos. Este artículo subraya desafíos técnicos, éticos y de privacidad, además de destacar el potencial de integrar IA e Internet de las Cosas (Internet of Things) (IoT) para fortalecer la ciberseguridad.

2.2.2 Marcos de Referencia en Ciberseguridad

El artículo *Artificial Intelligence for Cybersecurity: Literature Review and Future Research Directions* Ramanpreet Kaur (2023) destaca cómo la integración de la IA en la ciberseguridad se organiza eficazmente mediante marcos de referencia como el NIST, que agrupa las funciones clave en categorías que abordan desde la identificación hasta la recuperación frente a ciberataques. Según este enfoque, la gestión de activos, el análisis de riesgos y el modelado de amenazas son esenciales para una correcta identificación de riesgos. Por otro lado, la protección se centra en aspectos como la autenticación y la formación en ciberseguridad, mientras que la detección de intrusiones y malware juega un papel fundamental en la fase de monitoreo.

Además, este marco enfatiza la importancia de la automatización en la respuesta a incidentes, un área donde la IA ha mostrado un gran potencial para minimizar el impacto de los ataques en tiempo real. Finalmente, la resiliencia se aborda desde la evaluación de impactos hasta el diseño de estrategias de recuperación robustas. A pesar de estas contribuciones, el artículo señala desafíos importantes, como la falta de estándares uniformes y la necesidad de métodos avanzados para mitigar amenazas emergentes, subrayando la urgencia de armonizar la tecnología con las normativas existentes.

2.2.3 IA y Computación Cuántica en Ciberseguridad

En el ámbito de la ciberseguridad, la combinación de IA y computación cuántica representa una de las líneas de investigación más prometedoras. El artículo *Enhancing Cyber Security Using Quantum Computing and Artificial Intelligence: A Review* Singh (2024) explora esta convergencia, destacando aplicaciones que van desde la criptografía avanzada hasta la detección y mitigación de amenazas en tiempo real. Por ejemplo, los generadores cuánticos de números aleatorios se perfilan como una herramienta clave para desarrollar sistemas criptográficos más seguros y resistentes frente a ataques futuros, incluidos aquellos basados en computación cuántica.

Asimismo, la IA complementa estas capacidades al analizar grandes volúmenes de datos en tiempos reducidos, identificando patrones que permiten optimizar redes y detectar amenazas complejas. Entre las aplicaciones más innovadoras, se encuentran los protocolos de comunicación segura basados en Distribución de Claves Cuánticas (Quantum Key Distribution) (QKD)¹, los cuales garantizan la inviolabilidad de las comunicaciones incluso frente a ataques de alta sofisticación. Sin embargo, la implementación de estas tecnologías enfrenta obstáculos significativos, como los elevados costos de desarrollo y la falta de estándares regulatorios que favorezcan su adopción masiva.

2.2.4 IA en Criptografía

El uso de la IA en criptografía ha transformado la manera en que se diseñan y evalúan los sistemas de cifrado modernos. Según el artículo *Advances and Challenges in Cryptography Using Artificial Intelligence* Grewal (2023), la IA juega un papel central en tareas como la detección de patrones complejos y la autenticación avanzada. Por ejemplo, técnicas de

¹Quantum Key Distribution: técnica que permite compartir claves de cifrado de forma segura usando principios cuánticos.

aprendizaje automático han demostrado ser eficaces para analizar grandes volúmenes de datos cifrados, identificando anomalías y optimizando la seguridad de los sistemas.

Entre los beneficios clave de la IA en criptografía se incluyen la capacidad de adaptación a nuevas amenazas en tiempo real y la automatización de procesos para detectar vulnerabilidades de forma proactiva. Esto ha permitido, además, reducir significativamente los tiempos de respuesta frente a incidentes. Sin embargo, persisten retos importantes, como la falta de explicabilidad en los algoritmos utilizados, lo que dificulta la confianza en decisiones críticas tomadas por sistemas que actúan como "cajas negras". Además, el entrenamiento y despliegue de modelos avanzados suelen requerir recursos computacionales intensivos, lo que limita su aplicación en entornos con restricciones de hardware.

A futuro, se identifican varias direcciones prometedoras, como el desarrollo de algoritmos de Inteligencia Artificial Explicable (Explainable Artificial Intelligence) (XAI) que mejoren la transparencia y confianza en los modelos criptográficos, así como la incorporación de IA en sistemas resistentes a la computación cuántica. También se plantea la creación de sistemas híbridos humano-IA que combinen la experiencia humana con la potencia de los modelos de IA, junto con estándares éticos y normativas claras para garantizar el uso responsable de estas tecnologías.

2.2.4.1 Criptografía de Texto con Aprendizaje Profundo

El artículo titulado *A Survey of Cryptographic Algorithms with Deep Learning* Naser (2023) analiza cómo los algoritmos de aprendizaje profundo pueden combinarse con técnicas criptográficas tradicionales para mejorar tanto la eficiencia como la seguridad en el cifrado de texto plano. Una de las principales contribuciones de este enfoque es la integración de métodos clásicos de cifrado con redes neuronales profundas, lo que permite optimizar los procesos de encriptación y desencriptación.

Entre las aplicaciones más relevantes, se incluye el uso de aprendizaje profundo para proteger texto confidencial durante su transmisión, asegurando tanto su integridad como su confidencialidad. Además, se destaca la optimización de funciones criptográficas que reduce el tiempo de procesamiento y el consumo de recursos, un beneficio clave en dispositivos con limitaciones de memoria y energía.

Sin embargo, este enfoque no está exento de desafíos. La implementación inicial de algoritmos basados en aprendizaje profundo requiere una cantidad significativa de recursos computacionales, lo que puede ser una barrera en ciertos contextos. Asimismo, la generalización de estos modelos a diferentes tipos de texto y formatos representa un obstáculo que necesita ser superado. A pesar de estas limitaciones, el artículo concluye que el aprendizaje profundo tiene el potencial de revolucionar la criptografía de texto, especialmente si se avanza en la investigación hacia modelos que equilibren velocidad, eficiencia y robustez criptográfica.

2.2.4.2 Aplicaciones de Redes Neuronales en Criptografía

El uso de ANN en criptografía ha mostrado resultados prometedores en el fortalecimiento de la seguridad de sistemas criptográficos. Los artículos *Applications of Neural Network-Based AI in Cryptography* Abderrahmane Nitaj (2023) y *Neural Networks-Based Cryptography: A Survey* Sakurai (2021) destacan cómo las redes neuronales han sido utilizadas para optimizar

técnicas criptográficas, incluyendo redes adversariales generativas (GANs), esteganografía y esquemas criptográficos avanzados.

Históricamente, los primeros intentos, como el protocolo *Tree Parity Machine* desarrollado en los años 2000, expusieron vulnerabilidades significativas, pero allanaron el camino para desarrollos más avanzados. Desde 2016, las ANNs han permitido la creación de esquemas criptográficos más dinámicos y seguros, capaces de cifrar y descifrar datos sin depender de algoritmos predefinidos. Entre los algoritmos más destacados se encuentran Estándar de Encriptación Avanzado (Advanced Encryption Standard) (AES), Rivest-Shamir-Adleman (RSA), Aprendizaje con Errores (Learning With Errors) (LWE)² (*Learning With Errors*) y Ascon³.

En el caso del algoritmo AES, las ANNs se han utilizado para evaluar la resistencia de las S-boxes frente a ataques lineales y diferenciales, modelando propiedades críticas como la no linealidad y la uniformidad diferencial para optimizar su diseño. Por otro lado, en RSA, las ANNs han demostrado su utilidad para analizar la resistencia ante ataques de canal lateral y factoración, así como para generar claves públicas y privadas más seguras.

En el ámbito de LWE, las redes neuronales han permitido identificar parámetros vulnerables en problemas matemáticos basados en la teoría de retículas, reforzando la seguridad. Finalmente, en esquemas criptográficos ligeros como Ascon, las ANNs han sido empleadas para detectar vulnerabilidades frente a ataques conocidos, mejorando la robustez de las transformaciones criptográficas.

Adicionalmente, las ANNs han sido utilizadas en aplicaciones innovadoras como la esteganografía basada en GANs, protegiendo la privacidad de los datos y asegurando la integridad de la información. Asimismo, las ANNs se han integrado en esquemas post-cuánticos, diseñados para resistir las amenazas futuras derivadas de la computación cuántica.

A pesar de los beneficios, como la versatilidad en aplicaciones distribuidas y la capacidad de identificar vulnerabilidades proactivamente, las ANNs también enfrentan desafíos significativos. Por ejemplo, el ruido en los datos de entrenamiento puede afectar su precisión, mientras que la propensión al sobreajuste limita su capacidad de generalización. Además, la falta de interpretabilidad en los resultados dificulta la confianza en estas soluciones. Finalmente, los altos costos computacionales, tanto en entrenamiento como en implementación, representan una barrera para su adopción generalizada.

A pesar de estas limitaciones, las ANNs tienen un gran potencial para revolucionar la criptografía. La investigación actual se orienta hacia el desarrollo de generadores automáticos de redes neuronales, herramientas explicables (XAI) y la optimización de su integración en aplicaciones específicas. Estas innovaciones prometen un avance significativo en la creación de sistemas criptográficos robustos y adaptados a un entorno digital cada vez más complejo.

2.2.4.3 Aplicaciones Multidisciplinarias de IA en Criptografía

El artículo *Applications of Artificial Intelligence to Cryptography* Jonathan Blackledge (2020) presenta un enfoque multidisciplinario para el diseño y análisis de algoritmos criptográficos, combinando Aprendizaje Automático (Machine Learning) (ML), ANN y Computación

²Learning With Errors: problema matemático usado como base para esquemas criptográficos resistentes a ataques cuánticos.

³Algoritmo de cifrado ligero aprobado por NIST para dispositivos con recursos limitados.

Evolutiva (Evolutionary Computation) (EC). Este enfoque tiene el potencial de revolucionar la criptografía al integrar diversas técnicas avanzadas de IA.

Por un lado, el aprendizaje automático se utiliza para analizar y generar algoritmos de cifrado seguros, así como para realizar tareas de criptoanálisis que evalúan la aleatoriedad y detectan patrones en flujos de datos binarios. Por otro lado, las redes neuronales permiten la creación de códigos únicos e inclonables, mientras que la computación evolutiva contribuye al diseño de algoritmos criptográficos mediante la generación aleatoria de códigos y la evaluación de su robustez frente a ataques.

En términos de aplicaciones prácticas, el artículo destaca el uso de antenas de teléfonos inteligentes y etiquetas Identificación por Radiofrecuencia (Radio Frequency Identification) (RFID) flexibles para autenticar documentos de alto valor. También explora métodos ópticos basados en teléfonos inteligentes para descifrar códigos visuales y verificar documentos, lo que representa una innovación importante en autenticación y protección de datos.

A pesar de estas contribuciones, se identifican retos, como la necesidad de optimizar el equilibrio entre seguridad y eficiencia computacional, así como de adaptar las técnicas a entornos distribuidos y dispositivos con recursos limitados. Este enfoque multidisciplinario demuestra cómo la sinergia entre diferentes técnicas de IA puede transformar la criptografía y fortalecer la seguridad en un mundo digital cada vez más interconectado.

2.2.4.4 Sistemas Avanzados de Cifrado con Redes Neuronales y Sistemas Caóticos

El artículo *Cryptography Using Artificial Intelligence* Arpita Gupta (2020) explora el uso combinado de redes neuronales artificiales (ANNs) y sistemas caóticos para superar las limitaciones de los algoritmos criptográficos tradicionales. Este enfoque innovador se centra en la generación de claves criptográficas avanzadas, el cifrado eficiente de datos y la mejora de la seguridad frente a ataques predictivos.

Una de las principales contribuciones de este trabajo es el uso de redes neuronales para optimizar la generación de claves criptográficas, reduciendo el tiempo de entrenamiento y mejorando el rendimiento general del sistema. Además, se implementan algoritmos de retroalimentación que garantizan mayor precisión y eficiencia en los procesos de cifrado y descifrado.

El sistema también integra redes neuronales caóticas y máquinas de estados secuenciales, lo que asegura una alta aleatoriedad y sensibilidad frente a ataques. Esta combinación depende de parámetros iniciales específicos, lo que incrementa significativamente la dificultad de descifrar los datos sin las claves correctas. Entre las aplicaciones destacadas, se incluye el cifrado de señales digitales, textos y datos binarios, utilizando redes neuronales caóticas y redes secuenciales como la de Michael I. Jordan Contributors (2025), que ofrecen una mayor seguridad.

En cuanto a los resultados experimentales, las simulaciones muestran que estas arquitecturas producen salidas cifradas con alta precisión y variabilidad, dependiendo de las condiciones iniciales. Además, el tiempo de entrenamiento y los recursos computacionales requeridos se optimizan, haciendo que este enfoque sea más eficiente.

Sin embargo, el artículo también identifica desafíos importantes, como la complejidad en la implementación de sistemas caóticos y redes secuenciales en entornos distribuidos, así como los elevados requerimientos computacionales durante el entrenamiento inicial. A pesar de estas limitaciones, este enfoque demuestra cómo las ANNs combinadas con sistemas caóticos

pueden ofrecer soluciones innovadoras y eficientes para mejorar la seguridad criptográfica, abriendo nuevas posibilidades para el desarrollo de algoritmos de cifrado avanzados.

Además del papel de la IA en criptografía, otra aplicación disruptiva es el uso de redes generativas para reforzar la seguridad desde un enfoque sintético y predictivo.

2.3 Redes Adversariales Generativas (GAN) en Ciberseguridad

Las redes adversariales generativas (GAN) han emergido como una herramienta revolucionaria dentro del ámbito de la ciberseguridad, gracias a su capacidad para generar datos sintéticos y abordar desafíos complejos en la detección y prevención de amenazas. Estas redes, formadas por dos modelos que compiten entre sí—el generador y el discriminador—han mostrado resultados prometedores en diversos contextos.

2.3.1 Aplicaciones de las GANs

El uso de GANs para fortalecer los sistemas de ciberseguridad ha sido destacado en múltiples estudios. Según el artículo *The Role of Artificial Intelligence in Cybersecurity: Automation of Protection and Detection of Threats* Lysenko (2024), las GANs son especialmente efectivas para generar datos sintéticos que fortalecen la capacidad de los sistemas para identificar amenazas complejas y dinámicas. Este enfoque permite entrenar modelos de aprendizaje automático con datos simulados de alta calidad, mejorando su precisión y robustez frente a escenarios adversos.

Por otro lado, *A Survey of Cryptographic Algorithms with Deep Learning* Naser (2023) analiza cómo las GANs pueden integrarse en sistemas criptográficos para aumentar su resistencia frente a ataques. Su capacidad para generar datos realistas es aprovechada en pruebas de seguridad, simulando posibles vulnerabilidades en entornos controlados. Esto no solo mejora la resiliencia de los algoritmos criptográficos, sino que también acelera el desarrollo de nuevas soluciones más seguras.

2.3.2 Explicabilidad y Transparencia en las GANs

Uno de los retos asociados al uso de GANs en ciberseguridad es la necesidad de mejorar su explicabilidad y transparencia. En este sentido, el artículo *A Survey on Explainable Artificial Intelligence for Cybersecurity* Cohen (2023) examina cómo las GANs pueden contribuir a hacer más interpretables los modelos de IA. Entre las herramientas propuestas se incluyen modelos visuales que representan gráficamente las decisiones de las GANs, facilitando su comprensión para los desarrolladores y analistas de seguridad.

Adicionalmente, el artículo destaca el uso de análisis causal para identificar relaciones causa-efecto en los patrones de amenazas detectados. Estas técnicas no solo mejoran la interpretabilidad de las decisiones de las GANs, sino que también ayudan a identificar puntos débiles en los sistemas. Sin embargo, persisten desafíos significativos, como la alta complejidad de los modelos y los elevados costos computacionales asociados con la implementación de herramientas explicables.

2.4 Limitaciones y Oportunidades Futuras

A pesar de los avances significativos logrados con las GANs en el campo de la ciberseguridad, su integración enfrenta diversos desafíos. Entre las limitaciones más destacadas se encuentran la escalabilidad y la robustez de los modelos, que deben adaptarse a un entorno en constante cambio. Además, los costos computacionales elevados representan un obstáculo para su implementación masiva, especialmente en sistemas con recursos limitados.

Por otra parte, la falta de estándares regulatorios y éticos dificulta la adopción de estas tecnologías en sectores críticos, mientras que la complejidad de los modelos limita su aplicabilidad en entornos que requieren decisiones rápidas y confiables.

A pesar de estas limitaciones, las oportunidades son igualmente prometedoras. Por ejemplo, se vislumbra un futuro donde las GANs se integren de manera más efectiva mediante marcos como el NIST, lo que facilitará la armonización entre tecnología y normativas. Asimismo, se están desarrollando enfoques para mejorar la explicabilidad de las GANs, permitiendo la creación de sistemas más robustos y confiables. Finalmente, su capacidad para abordar nuevas amenazas emergentes posiciona a las GANs como una herramienta clave en la ciberseguridad de entornos digitales cada vez más complejos.

2.5 Conclusiones del Estado del Arte

La revisión del estado del arte subraya cómo la IA y, específicamente, las GANs están transformando el panorama de la ciberseguridad. Estas tecnologías ofrecen soluciones innovadoras para enfrentar desafíos emergentes, mejorando la capacidad de detección y prevención de amenazas en sistemas críticos. Este trabajo se enfocará en integrar estas innovaciones en el diseño de soluciones prácticas, particularmente en el ámbito de la seguridad para sistemas de reconocimiento facial biométrico, demostrando su potencial para reforzar la confianza en infraestructuras digitales.

3 Fundamentos Teóricos

3.1 Introducción

En la última década, la convergencia entre la IA y la ciberseguridad ha transformado significativamente el panorama tecnológico. La IA ha permitido el desarrollo de herramientas avanzadas para la detección, prevención y respuesta ante amenazas cibernéticas, facilitando el análisis de grandes volúmenes de datos en tiempo real y mejorando la eficiencia de los sistemas de seguridad Ayerbe (2020). Sin embargo, esta misma tecnología también ha sido explotada por ciberdelincuentes para sofisticar sus ataques, lo que plantea desafíos emergentes en la protección de infraestructuras críticas y datos sensibles “Riesgos de IA y Ciberseguridad” (2025).

Dentro de este contexto, las GANs han adquirido un papel crucial en la ciberseguridad. Originalmente diseñadas para la generación de contenido sintético, estas redes han demostrado tanto aplicaciones defensivas como ofensivas en el ámbito de la seguridad informática. Por un lado, pueden utilizarse para fortalecer modelos de detección de amenazas mediante la generación de datos de entrenamiento realistas; por otro, han sido empleadas para la creación de deepfakes y la evasión de mecanismos de autenticación biométrica “Los usos criminales de los ‘deepfakes’ se disparan: estafas, pornografía y suplantación de identidad” (2024).

Este capítulo establece las bases teóricas necesarias para comprender la relación entre IA y ciberseguridad, haciendo especial énfasis en el papel de las GANs en este entorno. Se abordarán los conceptos clave de ciberseguridad, los principios de aprendizaje automático aplicados a la detección de amenazas y la forma en que las GANs han revolucionado tanto la ofensiva como la defensiva en el ciberespacio “Generative Adversarial Networks: Inteligencia Artificial & Ciberseguridad (1 de 2)” (2018).

3.2 Conceptos Fundamentales en Ciberseguridad

3.2.1 Definición de Ciberseguridad

La ciberseguridad se refiere al conjunto de tecnologías, prácticas y políticas diseñadas para proteger sistemas informáticos, redes, dispositivos y datos contra ataques cibernéticos o accesos no autorizados. Su objetivo principal es salvaguardar la integridad, confidencialidad y disponibilidad de la información, previniendo amenazas como malware, phishing, ransomware y otras formas de ciberataques IBM (2024).

En el contexto actual, donde la digitalización es omnipresente, la ciberseguridad es esencial para garantizar la protección de datos personales y corporativos, así como para mantener la confianza en las interacciones digitales. La creciente sofisticación de los ciberataques y la dependencia de las infraestructuras tecnológicas hacen que la ciberseguridad sea un componente crítico en la gestión de riesgos de cualquier organización Pirani (2024).

3.2.1.1 Importancia de la Ciberseguridad

La ciberseguridad es vital para prevenir el acceso no autorizado a información sensible, proteger activos financieros y propiedad intelectual, y asegurar la continuidad operativa de las organizaciones. Además, una sólida postura de ciberseguridad ayuda a mantener la reputación de una empresa y garantiza el cumplimiento de normativas y regulaciones vigentes Elastic (2024).

La implementación de medidas de ciberseguridad no solo protege contra pérdidas financieras y de datos, sino que también fortalece la resiliencia frente a posibles interrupciones operativas y salvaguarda la confianza de clientes y socios comerciales en el entorno digital ENAE International Business School (2024).

3.2.2 Principales Amenazas y Vulnerabilidades

Las amenazas y vulnerabilidades en ciberseguridad representan un desafío constante en la protección de la información y la infraestructura digital. A medida que las tecnologías evolucionan, también lo hacen las técnicas utilizadas por ciberdelincuentes para comprometer la seguridad de los sistemas. A continuación, se describen algunas de las principales amenazas y su impacto en sistemas de autenticación biométrica.

3.2.2.1 Malware

El malware es un software malicioso diseñado para infiltrarse en un sistema sin el conocimiento del usuario, con el objetivo de causar daño, robar información o alterar su funcionamiento IBM (2024). Entre las principales variantes de malware se incluyen:

- **Virus:** Se adjuntan a archivos legítimos y se replican al ejecutarse.
- **Gusanos:** Se propagan automáticamente sin necesidad de intervención humana.
- **Troyanos:** Se disfrazan como software legítimo para obtener acceso no autorizado.
- **Spyware:** Monitorea la actividad del usuario sin su consentimiento.

3.2.2.2 Ransomware

El ransomware es una variante de malware que cifra los archivos de la víctima y exige un rescate para restaurar el acceso Cibersafety (2024). Este tipo de ataque ha aumentado considerablemente en los últimos años, afectando tanto a usuarios individuales como a empresas y organismos gubernamentales. En algunos casos, los atacantes amenazan con divulgar información sensible si no se paga el rescate.

3.2.2.3 Phishing

El phishing es una técnica de ingeniería social utilizada para engañar a las víctimas y obtener credenciales de acceso, información bancaria o datos personales CISA (2025). Los ciberdelincuentes suelen utilizar correos electrónicos, mensajes de texto o sitios web falsos que imitan servicios legítimos para persuadir a los usuarios de que revelen su información.

3.2.2.4 Ataques Adversariales

En el contexto de la IA, los ataques adversariales consisten en la manipulación intencional de entradas a un modelo de aprendizaje automático para engañarlo ArticAI (2025). Por ejemplo, pequeñas modificaciones en una imagen pueden hacer que un sistema de reconocimiento facial clasifique erróneamente una identidad, lo que representa un grave riesgo en sistemas biométricos de autenticación.

3.2.2.5 Impacto en Sistemas de Autenticación Biométrica

Los sistemas biométricos, que utilizan características físicas o conductuales para la autenticación, son susceptibles a diversas amenazas:

- **Suplantación de identidad:** Uso de huellas dactilares falsas, fotografías o videos para engañar sistemas de autenticación.
- **Deepfakes:** Utilización de GANs para generar rostros sintéticos que burlen sistemas de reconocimiento facial Directores de Seguridad (2024).
- **Ataques de inyección de datos:** Modificación de entradas para forzar respuestas incorrectas en sistemas de verificación.

Ante estas amenazas, es fundamental implementar medidas de seguridad avanzadas, como la detección de ataques adversariales, el uso de autenticación multifactor y la monitorización continua de accesos para mitigar los riesgos en sistemas biométricos.

3.2.3 Modelos de Seguridad en Sistemas de Información

En el ámbito de la ciberseguridad, los modelos de seguridad proporcionan marcos conceptuales que ayudan a las organizaciones a proteger sus activos de información. Entre los más reconocidos se encuentra la tríada CIA y el modelo de Confianza Cero (Zero Trust), complementados por estrategias como la defensa en profundidad.

3.2.3.1 Tríada CIA: Confidencialidad, Integridad y Disponibilidad

La tríada CIA es un modelo fundamental en la seguridad de la información que se centra en tres pilares esenciales:

- **Confidencialidad:** Garantiza que la información sea accesible únicamente por personas autorizadas, protegiendo los datos sensibles de accesos no autorizados. Esto se logra mediante técnicas como el cifrado y controles de acceso estrictos Fortinet (2024).
- **Integridad:** Asegura que la información se mantenga precisa y completa, evitando modificaciones no autorizadas. Mecanismos como las sumas de verificación (checksums) y las firmas digitales son comunes para preservar la integridad de los datos DataSunrise (2024).
- **Disponibilidad:** Se enfoca en que los sistemas y datos estén disponibles para los usuarios autorizados cuando los necesiten, implementando medidas como redundancia, planes de recuperación ante desastres y mantenimiento regular de sistemas Comillas (2024).

3.2.3.2 Modelo de Confianza Cero (Zero Trust)

El modelo de Confianza Cero es una estrategia de seguridad que opera bajo el principio de "nunca confiar, siempre verificar". A diferencia de los enfoques tradicionales que asumían confianza implícita dentro del perímetro de la red, este modelo requiere una autenticación y autorización estrictas para cada acceso a recursos, independientemente de su origen. Los componentes clave incluyen:

- **Autenticación y autorización continuas:** Verificación constante de la identidad y permisos de usuarios y dispositivos antes de otorgar acceso Akamai (2024).
- **Acceso con privilegios mínimos:** Los usuarios reciben únicamente los permisos necesarios para cumplir sus funciones, minimizando el riesgo de accesos no autorizados First (2024).
- **Segmentación de la red:** División de la red en segmentos más pequeños para limitar el movimiento lateral de posibles atacantes Cloudflare (2024).

3.2.3.3 Defensa en Profundidad

La defensa en profundidad es una estrategia que implementa múltiples capas de seguridad para proteger los recursos de una organización. Este enfoque asegura que si una capa de defensa es comprometida, las capas adicionales continúen proporcionando protección. Las capas pueden incluir:

- **Controles de seguridad físicos:** Como sistemas de autenticación biométrica y vigilancia.
- **Controles técnicos:** Incluyen *firewalls* (sistemas de filtrado de red), sistemas de detección de intrusiones y cifrado de datos.
- **Controles administrativos:** Políticas y procedimientos de seguridad, junto con programas de concienciación y formación para empleados Cloudflare (2024).

La integración de estos modelos y estrategias permite a las organizaciones establecer una postura de seguridad robusta, adaptándose a las amenazas emergentes y protegiendo eficazmente sus activos de información.

3.2.4 Marcos Regulatorios y Normativas de Seguridad

En el ámbito de la ciberseguridad, existen diversas regulaciones y normativas que establecen directrices para proteger la información y garantizar la privacidad. A continuación, se describen algunas de las más relevantes y su impacto en la implementación de la IA en ciberseguridad.

3.2.4.1 Reglamento General de Protección de Datos (GDPR)

El GDPR es una regulación de la Unión Europea que entró en vigor en 2018, centrada en la protección de datos personales y la privacidad de los individuos. Establece obligaciones estrictas para las organizaciones en cuanto a la recopilación, almacenamiento y procesamiento de datos personales. En el contexto de la IA aplicada a la ciberseguridad, el GDPR impone restricciones sobre cómo se pueden utilizar los datos personales para entrenar modelos de IA, asegurando que se respeten los derechos de los individuos y se mantenga la transparencia en el procesamiento de datos.

3.2.4.2 ISO/IEC 27001

La norma internacional **ISO/IEC 27001** proporciona un marco para establecer, implementar, mantener y mejorar continuamente un Sistema de Gestión de Seguridad de la Información (SGSI). Esta norma ayuda a las organizaciones a proteger sus activos de información mediante la implementación de controles de seguridad adecuados. Al integrar soluciones de IA en ciberseguridad, es esencial que estas cumplan con los controles y procesos establecidos por la ISO/IEC 27001 para garantizar la confidencialidad, integridad y disponibilidad de la información.

3.2.4.3 Marco de Ciberseguridad del NIST

El Marco de Ciberseguridad del NIST es una guía desarrollada por esta entidad, que ofrece recomendaciones para gestionar y reducir los riesgos de ciberseguridad. Aunque fue diseñado inicialmente para organizaciones estadounidenses, su flexibilidad permite que sea adoptado globalmente. La versión 2.0 del marco incorpora consideraciones específicas para la gestión de riesgos asociados con la IA, proporcionando directrices sobre cómo integrar sistemas de IA de manera segura y ética en las operaciones de ciberseguridad Núñez (2024).

3.2.4.4 Impacto en la Implementación de IA en Ciberseguridad

La implementación de soluciones de IA en ciberseguridad debe alinearse con las regulaciones y normativas mencionadas. Esto implica:

- **Gestión de Riesgos:** Evaluar y mitigar los riesgos específicos asociados con la IA, considerando aspectos técnicos, éticos y de cumplimiento normativo Insights (2024).
- **Transparencia y Explicabilidad:** Asegurar que los modelos de IA sean transparentes y sus decisiones explicables, facilitando la auditoría y el cumplimiento de regulaciones como el GDPR.
- **Protección de Datos:** Implementar medidas que garanticen la privacidad y seguridad de los datos utilizados por los sistemas de IA, en conformidad con estándares como la ISO/IEC 27001.
- **Actualización Continua:** Mantenerse al día con las evoluciones en los marcos regulatorios y adaptar las soluciones de IA en consecuencia para asegurar el cumplimiento continuo.

Al adherirse a estos marcos regulatorios y normativas de seguridad, las organizaciones pueden implementar soluciones de IA en ciberseguridad de manera responsable, minimizando riesgos y fortaleciendo la confianza en sus sistemas.

3.3 Inteligencia Artificial y Ciberseguridad

3.3.1 Conceptos Básicos de IA y Aprendizaje Automático

La **Inteligencia Artificial (IA)** es un campo de la informática centrado en la creación de sistemas capaces de realizar tareas que normalmente requieren inteligencia humana, como el razonamiento, la toma de decisiones, el reconocimiento de patrones o el aprendizaje a partir de datos Russell y Norvig (2021).

Dentro de la IA, el ML es una de las ramas más relevantes. Se basa en la idea de que los sistemas pueden aprender de los datos y mejorar su rendimiento sin ser explícitamente programados para cada tarea específica. Este aprendizaje puede clasificarse en tres tipos principales:

- **Aprendizaje supervisado:** Se entrena al modelo con datos de entrada y sus correspondientes salidas deseadas, permitiendo que el sistema aprenda una función que relacione ambos. Por ejemplo, la detección de *spam* en correos electrónicos.
- **Aprendizaje no supervisado:** El sistema analiza datos no etiquetados para encontrar patrones o estructuras ocultas. Un ejemplo de este tipo de aprendizaje es la agrupación de usuarios según su comportamiento.
- **Aprendizaje por refuerzo:** Se basa en recompensas y penalizaciones. El agente aprende a través de la interacción con su entorno, maximizando una señal de recompensa. Un ejemplo son los algoritmos de defensa adaptativa frente a ciberataques Sutton y Barto (2018).

En el ámbito de la ciberseguridad, estos enfoques permiten desarrollar sistemas capaces de identificar amenazas emergentes, detectar anomalías en el tráfico de red o realizar análisis predictivos. Por ejemplo, los algoritmos de clasificación supervisada se utilizan en IDS, mientras que el aprendizaje no supervisado se emplea en la detección de patrones inusuales sin necesidad de ejemplos previos. El aprendizaje por refuerzo, aunque más reciente en este campo, está ganando interés en la automatización de respuestas ante incidentes de seguridad Alshamrani y Alkasassbeh (2021).

3.3.2 Uso de IA en Ciberseguridad

La incorporación de la IA en ciberseguridad ha abierto nuevas posibilidades para detectar y mitigar amenazas de manera proactiva. Gracias a su capacidad para analizar grandes volúmenes de datos en tiempo real, los sistemas basados en IA se han convertido en herramientas esenciales para la protección de infraestructuras digitales críticas.

Entre las aplicaciones más destacadas se encuentran:

- **Sistemas de detección de intrusos (IDS):** Utilizan algoritmos de aprendizaje automático para identificar comportamientos anómalos que podrían indicar accesos no autorizados o ataques en curso Sommer y Paxson (2010).
- **Análisis de tráfico de red:** Los modelos de IA permiten detectar patrones sospechosos en el tráfico, diferenciando entre actividades legítimas y posibles ataques, como escaneos de puertos o ataques de Ataque Distribuido de Denegación de Servicio (Distributed Denial of Service) Javaid y cols. (2016).
- **Prevención y detección de fraudes:** En el sector financiero, se emplean modelos supervisados y no supervisados para identificar operaciones fraudulentas basadas en desviaciones de patrones de comportamiento habituales Alabdallah y cols. (2016).
- **Automatización de respuestas:** Sistemas basados en IA que pueden actuar automáticamente al detectar amenazas, mitigando los efectos antes de que se produzcan daños mayores.
- **Análisis de malware:** La clasificación automática de archivos maliciosos mediante redes neuronales o árboles de decisión reduce el tiempo necesario para responder ante nuevos tipos de amenazas.

Estas aplicaciones no solo mejoran la eficiencia de los equipos de seguridad, sino que también reducen el tiempo de detección y respuesta ante incidentes, un factor clave en la prevención de daños significativos.

3.3.3 Ataques Adversariales en IA

Los ataques adversariales representan una amenaza emergente y crítica en los sistemas de IA, especialmente en aquellos que utilizan modelos de aprendizaje profundo. Estos ataques consisten en introducir pequeñas perturbaciones intencionadas en los datos de entrada — como imágenes, texto o señales— con el objetivo de engañar al modelo y provocar predicciones incorrectas, sin que estas alteraciones sean perceptibles al ojo humano.

Este tipo de ataques puede comprometer gravemente la seguridad de sistemas automatizados en contextos como el reconocimiento facial, la conducción autónoma o la detección de amenazas en ciberseguridad. En un entorno donde los sistemas de defensa se basan en IA para detectar intrusiones o clasificar contenido malicioso, la capacidad de un atacante para generar ejemplos adversariales que pasen desapercibidos puede reducir significativamente la eficacia del sistema Akhtar y Mian (2018).

Entre las técnicas de ataque adversarial más estudiadas se encuentran:

- **Método de Signo de Gradiente Rápido (Fast Gradient Sign Method) (FGSM):** Propuesto por I. J. Goodfellow y cols. (2015), este método fue uno de los primeros en demostrar lo vulnerables que pueden ser los modelos de redes neuronales ante pequeñas alteraciones. El FGSM funciona calculando el gradiente de la función de pérdida del modelo con respecto a la entrada original, y luego modifica la entrada en la dirección de ese gradiente, con una magnitud controlada por un parámetro ϵ . La nueva entrada, aunque muy similar a la original, puede inducir al modelo a cometer un error. Es

un método rápido y eficaz que pone de manifiesto la sensibilidad de los modelos ante cambios mínimos.

- **DeepFool:** Esta técnica, introducida por Moosavi-Dezfooli y cols. (2016), tiene como objetivo encontrar la mínima perturbación posible que consiga cambiar la clasificación de una entrada por parte del modelo. A diferencia de FGSM, que aplica una única alteración basada en el gradiente, DeepFool realiza un proceso iterativo de aproximación lineal al clasificador para encontrar el punto más cercano en el que cambia la decisión del modelo. De este modo, produce ejemplos adversariales más sutiles y difíciles de detectar, siendo útil para evaluar la verdadera robustez de un sistema .
- **Carlini & Wagner (C&W):** Este método, desarrollado por Carlini y Wagner (2017), se considera uno de los ataques más eficaces y sofisticados contra redes neuronales profundas. Su objetivo es generar perturbaciones que sean imperceptibles para el ojo humano, utilizando funciones de pérdida cuidadosamente diseñadas y técnicas de optimización más avanzadas. A diferencia de FGSM y DeepFool, que se centran en la velocidad o la simplicidad, C&W prioriza la invisibilidad del ataque y su capacidad para evadir incluso mecanismos defensivos. Este ataque ha sido ampliamente utilizado para probar la seguridad de modelos en tareas críticas como la clasificación de imágenes o la autenticación biométrica.

La existencia de estos ataques ha dado lugar al desarrollo de estrategias de defensa como la regularización adversarial, la detección de entradas anómalas o el entrenamiento robusto. No obstante, los atacantes continúan desarrollando técnicas más sofisticadas, por lo que la defensa frente a ejemplos adversariales sigue siendo un área activa de investigación en ciberseguridad e IA.

3.4 Redes Adversariales Generativas (GANs)

3.4.1 Concepto y Arquitectura de las GANs

Las GANs son un tipo de modelo de aprendizaje profundo introducido por I. J. Goodfellow y cols. (2014). Su propósito principal es generar datos sintéticos que imiten con gran realismo una distribución de datos reales, lo que ha abierto nuevas posibilidades en múltiples campos, incluyendo la ciberseguridad.

Las GANs constan de dos redes neuronales que compiten entre sí en un proceso de aprendizaje conjunto: un generador y un discriminador. Esta configuración se basa en un juego de suma cero, en el que el generador busca crear muestras realistas para “engaños” al discriminador, mientras que el discriminador intenta identificar correctamente si una muestra es real o generada.

Estructura de las GANs

- **Generador (G):** Toma como entrada un vector de ruido aleatorio, proveniente de una distribución latente (por ejemplo, gaussiana), y lo transforma en una muestra sintética. Utiliza capas densas o convolucionales, activaciones no lineales y técnicas de normalización para producir datos que simulen las características de los datos reales.

- **Discriminador (D):** Recibe como entrada tanto muestras reales como generadas y determina si una muestra proviene del conjunto de entrenamiento o ha sido creada por el generador. Para ello, emplea arquitecturas típicas de clasificación binaria, con funciones de activación como ReLU y sigmoide en la salida.

Proceso de Entrenamiento

El entrenamiento de una GANs es un proceso iterativo en el que ambas redes aprenden de manera simultánea. Primero se entrena el discriminador con un conjunto mixto de muestras reales y generadas. A continuación, se entrena el generador para que produzca muestras cada vez más convincentes, basándose en el feedback del discriminador. La función de pérdida utilizada refleja esta competencia y se actualiza gradualmente para equilibrar el rendimiento de ambos modelos Creswell y cols. (2018).

Una GANs alcanza un estado óptimo cuando el discriminador ya no puede diferenciar entre muestras reales y falsas con una precisión superior al azar (50%). En la práctica, el entrenamiento es inestable y requiere técnicas como el uso de *label smoothing*, penalización por gradiente o arquitectura Wasserstein Generative Adversarial Network (WGAN) para mejorar la convergencia.

Importancia de las GANs en la Generación de Datos Sintéticos

La capacidad de generar datos sintéticos realistas convierte a las GANs en herramientas valiosas para contextos donde los datos reales son escasos, costosos o sensibles, como en medicina, biometría o ciberseguridad. En estos entornos, las GANs permiten crear ejemplos adicionales para mejorar el entrenamiento de otros modelos, detectar vulnerabilidades mediante datos artificiales o incluso simular escenarios de ataque para robustecer la defensa de los sistemas Pan y cols. (2021); Hong y cols. (2022).

3.4.2 Aplicaciones de GANs en Ciberseguridad

Las GANs ofrecen un amplio abanico de aplicaciones en el campo de la ciberseguridad, tanto desde la perspectiva ofensiva como desde un enfoque defensivo. Su capacidad para generar datos altamente realistas, imitando patrones complejos, las convierte en herramientas útiles, pero también potencialmente peligrosas.

3.4.2.1 Usos ofensivos: ataques mediante GANs

En el contexto ofensivo, las GANs pueden utilizarse por actores maliciosos para crear contenido falsificado que supere controles automatizados de verificación. Algunas de las aplicaciones más relevantes incluyen:

- **Suplantación de identidad y deepfakes:** Las GANs pueden generar imágenes, voces o vídeos falsos de personas reales. Este fenómeno, conocido como *deepfake*, puede emplearse para suplantar identidades en videollamadas, ataques de ingeniería social o manipulación de sistemas biométricos Otto (2024).

- **Generación de malware polimórfico:** Aunque aún en etapas experimentales, se ha explorado el uso de GANs para crear variantes de malware que escapan a la detección de antivirus tradicionales al modificar su firma digital sin alterar su funcionalidad.

3.4.2.2 Usos defensivos: protección mediante GANs

A pesar de los riesgos, las GANs también pueden ser aliadas en la mejora de las defensas ciberneticas. Entre sus aplicaciones más destacadas se encuentran:

- **Detección de contenido falsificado:** Paradójicamente, las mismas técnicas que permiten generar *deepfakes* pueden utilizarse para entrenar modelos que los detecten. Al simular múltiples variantes de falsificación, los sistemas defensivos pueden anticiparse a nuevas amenazas y mejorar su capacidad de detección Greydon (2022).
- **Generación de datos sintéticos para entrenamiento:** En ciberseguridad, los conjuntos de datos reales pueden ser limitados o contener información sensible. Las GANs permiten crear datos sintéticos que reproducen las características estadísticas de los datos reales sin comprometer la privacidad. Estos datos pueden utilizarse para entrenar modelos de detección de intrusiones, clasificación de tráfico de red o análisis de vulnerabilidades Frid-Adar y cols. (2018).

Esta dualidad —como herramienta de ataque y defensa— convierte a las GANs en un tema crucial dentro del estudio de la IA aplicada a la ciberseguridad. Su estudio y regulación se vuelven esenciales para anticipar riesgos, al mismo tiempo que se explotan sus beneficios en contextos éticos y seguros.

3.4.3 Retos y Desafíos de las GANs en Seguridad

Las GANs han demostrado ser herramientas poderosas en el ámbito de la ciberseguridad. Sin embargo, su implementación presenta una serie de desafíos y limitaciones que deben ser considerados para garantizar su uso efectivo y responsable.

Explicabilidad y Transparencia

Uno de los principales desafíos en el uso de GANs es la falta de explicabilidad de sus modelos. Las GANs operan como *cajas negras*, lo que dificulta comprender cómo generan sus resultados. Esta opacidad puede ser problemática en ciberseguridad, donde es crucial entender el proceso de toma de decisiones para confiar en las herramientas y detectar posibles vulnerabilidades. La necesidad de técnicas que permitan interpretar y auditar el comportamiento de las GANs es, por tanto, evidente PricewaterhouseCoopers (2021).

Consideraciones Éticas

El uso de GANs plantea importantes cuestiones éticas. Su capacidad para generar contenido sintético realista puede ser mal utilizada para crear *deepfakes*, facilitando la desinformación o la suplantación de identidad. Además, la generación de datos sintéticos puede perpetuar sesgos existentes si los datos originales contienen prejuicios, afectando negativamente a grupos

específicos. Es esencial establecer directrices éticas claras para el desarrollo y aplicación de GANs, asegurando que su uso respete los derechos fundamentales y promueva la equidad González (2020).

Regulaciones y Cumplimiento Normativo

El marco regulatorio en torno a la IA, y en particular a las GANs, está en constante evolución. Legislaciones como el Reglamento de IA de la Unión Europea buscan establecer límites y clasificaciones de riesgo para los sistemas de IA, imponiendo obligaciones a proveedores y usuarios para garantizar la seguridad y protección de los ciudadanos País (2024). Estas regulaciones pueden limitar ciertas aplicaciones de las GANs, especialmente aquellas que impliquen riesgos elevados, y requieren que las organizaciones se mantengan actualizadas y cumplan con las normativas vigentes.

Si bien las GANs ofrecen oportunidades significativas en ciberseguridad, es fundamental abordar sus desafíos relacionados con la explicabilidad, ética y regulaciones. Solo mediante un enfoque consciente y responsable se podrá aprovechar su potencial sin comprometer la seguridad, la confianza y los valores fundamentales de la sociedad.

3.5 Síntesis y Relación con la Solución Propuesta

Este capítulo ha establecido los pilares conceptuales que sustentan el enfoque del presente trabajo en el ámbito de la ciberseguridad y la inteligencia artificial. Se han descrito los principios fundamentales de la ciberseguridad, los modelos y normativas que la regulan, así como las principales amenazas que enfrentan los sistemas de información, con especial atención a aquellos basados en autenticación biométrica.

Además, se han introducido los conceptos clave de la inteligencia artificial y el aprendizaje automático, mostrando su aplicabilidad tanto en la detección como en la mitigación de ciberataques. En particular, se ha analizado en profundidad el fenómeno de los ataques adversariales, que ponen de manifiesto la vulnerabilidad de los modelos de IA ante manipulaciones sútiles pero efectivas.

Dentro de este contexto, las GANs han sido presentadas como una tecnología con gran potencial tanto ofensivo como defensivo. Su capacidad para generar datos sintéticos realistas las convierte en herramientas poderosas que, si bien pueden emplearse para atacar sistemas de reconocimiento facial mediante técnicas como los *deepfakes*, también pueden ser aprovechadas para fortalecer dichos sistemas frente a intentos de suplantación.

La comprensión teórica de estos elementos será esencial para el desarrollo posterior del trabajo, donde se explorará cómo aplicar estas capacidades de las GANs —junto con análisis estructural mediante mapas de profundidad— a la mejora de los sistemas de autenticación biométrica. Este capítulo, por tanto, sirve como fundamento para justificar y contextualizar la solución que se planteará en los siguientes apartados.

4 Metodología

4.1 Enfoque Metodológico

El desarrollo de este Trabajo de Fin de Máster se ha sustentado en un enfoque metodológico estructurado, basado en la división del trabajo en dos líneas de investigación independientes, pero convergentes en sus objetivos. Ambas líneas han sido concebidas con el propósito de fortalecer la fiabilidad de los sistemas de verificación facial, especialmente frente a ataques que utilizan imágenes impresas, manipuladas o generadas de forma sintética.

Por un lado, se ha llevado a cabo el diseño y la implementación de un sistema basado en redes generativas adversariales (GANs), con el objetivo de explorar la generación de rostros sintéticos y el posterior entrenamiento de un discriminador capaz de diferenciarlos de imágenes reales. Esta línea se ha centrado en la construcción progresiva de un modelo GANs funcional, su entrenamiento con diferentes resoluciones y datasets, y la evaluación visual y métrica de los resultados obtenidos. Se ha hecho uso de arquitecturas secuenciales simples, adaptadas a rostros en escala de grises, y se ha profundizado en la dinámica generador-discriminador, clave en el aprendizaje adversarial.

De forma separada, se ha trabajado en una segunda línea metodológica basada en el análisis Estructural de las imágenes medias mapa de Profundidad. Para ello, se ha utilizado el Modelo Preentrenado Midas, Con el Que se Han Generado Mapas de Profundidad A Partir de Imágenes faciales Tanto Reales como Falsificadas. Posterista, Se Han Aplicado Técnicas de Segmentación, Extracción de ROI y Análisis Estadístico (Desviación Estándar, Rango de Valores, Textura), Con el Fin de Identificar Patrones Diferenciadores Entre Imágenes Auténticas y Copias.

Ambos enfoques han sido desarrollados de forma independiente para permitir la validación individual de sus respectivas hipótesis y funcionalidades. No obstante, se ha mantenido como objetivo a medio plazo la posibilidad de integrar ambos sistemas en una arquitectura híbrida más robusta. De esta manera, el sistema basado en GANs podría ser complementado con un análisis estructural adicional, aprovechando la riqueza de la información de profundidad que MiDaS es capaz de proporcionar.

Esta separación metodológica ha permitido abordar los retos técnicos específicos de cada línea con mayor eficacia, favoreciendo al mismo tiempo la escalabilidad del proyecto. Asimismo, se ha documentado cada fase del desarrollo de forma exhaustiva, garantizando la reproducibilidad del trabajo y sentando las bases para futuras extensiones o integraciones.

4.2 Desarrollo de un Sistema Basado en GANs

mista Secia Recoge El Desarrollo Completo del Sistema Basado en Redes Generativas Antagónicas (GANs), El Cual Ha Sido Enfocado Con El Objetivo de Generar Múltiples imágenes de Rostros humanos en escala de Grises. Este enfoque Permite Estudiar Las

Capacidades de Las GANs para Simular La Distribución de Un Conjuntos de Datos, Así Como analizar su utilidad dentro de Sistemas de Verificación Expiaciones faciales manipuladas.

El Sistema se Estructura en torno a dos ronda neuronales Atrenadas De Manera Conjunta y Competitiva: El Generador, Aparato de productir soja Uélegidad Singotas Un para para para par de ruido aleatorio, y el discriminador Evaluación De la de la Evaluación De la Evaluación. real. Ambos Modelo Han Sido implementados Utilizando Arquitecturas Segenciales Y Funciones de Activaciós Especies, prestando Atentacia Especial entrada y salida.

4.3 Desarrollo de un Sistema Basado en GANs

Esta sección recoge el desarrollo completo del sistema basado en redes generativas antagónicas (GANs), el cual ha sido concebido con el objetivo de generar imágenes sintéticas de rostros humanos en escala de grises. Este enfoque permite estudiar las capacidades de las GANs para simular la distribución de un conjunto de datos reales, así como analizar su utilidad dentro de sistemas de verificación facial expuestos a intentos de suplantación mediante imágenes generadas o manipuladas.

El sistema se estructura en torno a dos redes neuronales entrenadas de manera conjunta y competitiva: el generador, encargado de producir imágenes sintéticas a partir de ruido aleatorio, y el discriminador, cuya función es evaluar si una imagen corresponde a una muestra real o ha sido generada. Ambos modelos han sido implementados utilizando arquitecturas secuenciales y funciones de activación específicas, prestando especial atención a la estabilidad del entrenamiento y a la coherencia entre los datos de entrada y salida.

4.3.1 Diseño del Generador y Discriminador

El diseño del generador y del discriminador se ha basado en una arquitectura secuencial compuesta por capas densas, activaciones LeakyReLU y técnicas de normalización por lotes. Esta elección responde a la necesidad de mantener un equilibrio entre simplicidad estructural y estabilidad en el entrenamiento, especialmente en las primeras fases del desarrollo.

El generador recibe como entrada un vector de ruido de dimensión fija (`latent_dim = 100`), que representa una muestra aleatoria del espacio latente. A partir de este vector, la red aplica de forma progresiva varias transformaciones mediante capas `Dense` de 256, 512 y 1024 unidades respectivamente, combinadas con activaciones `LeakyReLU` ($\alpha = 0.2$) y `BatchNormalization` (`momentum = 0.8`). La capa de salida utiliza una activación `tanh` para limitar el rango de valores entre $[-1, 1]$, y se completa con una capa `Reshape` que transforma el vector resultante en una imagen con la dimensión deseada (por ejemplo, 28x28x1, 52x52x1 o 96x96x1).

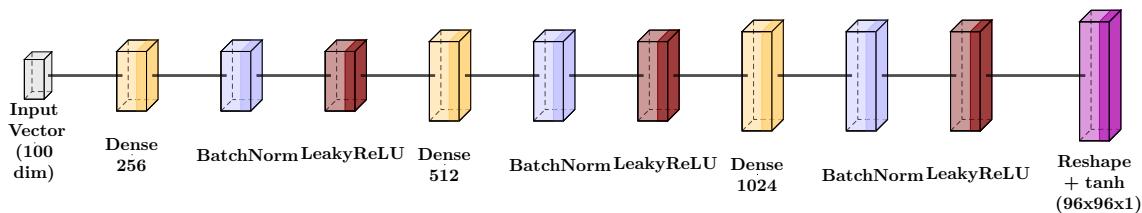


Figura 4.1: Arquitectura del generador basada en capas densas, normalización por lotes y activaciones LeakyReLU.

El discriminador, por su parte, está diseñado como un clasificador binario. Recibe como entrada una imagen en escala de grises con una forma determinada y la convierte en un vector unidimensional mediante una capa **Flatten**. A continuación, la información se procesa a través de dos capas **Dense** con 512 y 256 unidades, ambas con activación **LeakyReLU**. La salida se obtiene mediante una única neurona con activación **sigmoid**, que devuelve la probabilidad de que la imagen evaluada sea real.

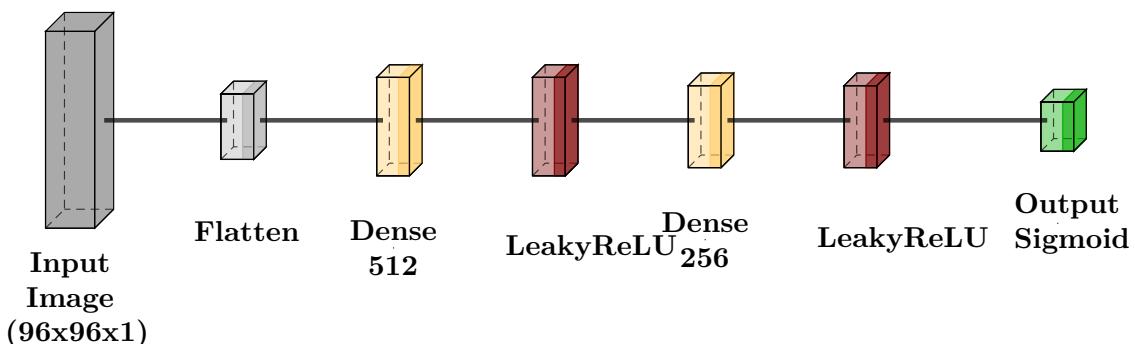


Figura 4.2: Arquitectura del discriminador basada en capas densas, activaciones LeakyReLU y salida sigmoid.

Todas las imágenes utilizadas han sido convertidas previamente a escala de grises, redimensionadas a la resolución objetivo y normalizadas al rango $[-1, 1]$, lo cual garantiza la compatibilidad con la salida del generador y con los requerimientos del discriminador. Esta coherencia en el preprocesamiento ha sido fundamental para facilitar la convergencia del modelo durante el entrenamiento y mejorar la calidad visual de las imágenes generadas.

En conjunto, la arquitectura definida permite estudiar de forma controlada el proceso de generación de imágenes sintéticas y ofrece una base sólida sobre la que construir experimentos más complejos en etapas posteriores del proyecto.

4.3.2 Preparación del Dataset

4.3.2.1 Preparación del conjunto de datos

El desarrollo del sistema ha requerido la utilización de distintos conjuntos de datos, adaptados a los objetivos específicos de cada etapa del proyecto. En particular, se han empleado tres datasets principales: uno de carácter experimental, otro destinado al entrenamiento de la red generativa y un tercero orientado a tareas de clasificación.

4.3.2.1.1 Conjunto de datos de validación experimental En la fase inicial del proyecto se ha utilizado el conjunto de datos MNIST, ampliamente reconocido en la literatura científica por su simplicidad y utilidad en entornos de validación. Este dataset contiene imágenes en escala de grises de dígitos manuscritos con una resolución de 28×28 píxeles. Su uso ha permitido verificar el correcto funcionamiento de la arquitectura de la red generativa, en un entorno controlado de baja complejidad, antes de abordar dominios visuales más exigentes como el de los rostros humanos.

En la Figura 4.3 se muestran algunos ejemplos de imágenes del conjunto MNIST, ilustrando la variedad de estilos y trazos presentes en los dígitos manuscritos.



Figura 4.3: Ejemplos de imágenes del dataset MNIST utilizadas para validación experimental.

4.3.2.1.2 Conjunto de datos para entrenamiento del generador Para entrenar el modelo generativo en un contexto realista, se ha utilizado un conjunto de datos compuesto por 21 vídeos de ataques por impresión, en los que se muestran rostros impresos en papel expuestos frente a una cámara. Estos vídeos, con una resolución de 1920×1080 píxeles, obtenidos a través de la plataforma Kaggle Kaggle (2024), han sido procesados para extraer un total de 100 fotogramas por archivo, seleccionando exclusivamente los 100 primeros fotogramas de cada vídeo. Esta selección tiene como objetivo representar adecuadamente la variabilidad del movimiento sin incurrir en una carga computacional excesiva.

Cada imagen ha sido convertida a escala de grises, redimensionada a 96×96 píxeles y normalizada al rango $[-1, 1]$, en consonancia con los requisitos de entrada de la red generativa. Las imágenes han sido organizadas en carpetas estructuradas dentro del directorio de trabajo del servidor. La siguiente figura muestra la estructura real del directorio utilizada para el entrenamiento:

```

tfm/fase3/
gan2.py
dataset/
    photo-print-attacks-dataset-1k-individuals.zip
    print_samples/
        print_1.mp4
        print_2.mp4
        ...
        print_21.mp4
generated_images/
    epoch_0.png
    epoch_100.png
    epoch_200.png
    ...
    epoch_900.png
best_generator_weights.h5
best_discriminator_weights.h5
generator_model.keras
discriminator_model.keras
combined_model.keras

```

Figura 4.4: Estructura real del directorio de trabajo para la versión de la GAN entrenada con el conjunto de datos de ataques por impresión obtenido de Kaggle.

En la Figura 4.5, Figura 4.6 y Figura 4.7 se muestran ejemplos de secuencias de fotogramas extraídos de tres vídeos distintos del conjunto de entrenamiento. Cada uno de los 21 vídeos utilizados en este conjunto corresponde a una persona diferente, lo que garantiza una adecuada diversidad de sujetos. Se puede observar cómo pequeñas variaciones en la perspectiva y el encuadre enriquecen la variedad visual de los datos generados, favoreciendo la robustez del modelo ante diferencias sutiles en las condiciones de captura.



Figura 4.5: Secuencia de frames extraídos del vídeo Print01.



Figura 4.6: Secuencia de frames extraídos del vídeo Print10.



Figura 4.7: Secuencia de frames extraídos del vídeo Print20.

4.3.2.1.3 Conjunto de datos para clasificación de emociones En paralelo al desarrollo del modelo generativo, se ha empleado el conjunto de datos FER2013 I. Goodfellow y cols. (2013), orientado a la clasificación de emociones a partir de expresiones faciales. Este dataset incluye un total de 35,887 imágenes en escala de grises, cada una con una resolución de 48×48 píxeles, capturadas en formato de primeros planos faciales. Las imágenes están etiquetadas en siete categorías emocionales: *angry*, *disgust*, *fear*, *happy*, *neutral*, *sad* y *surprise*.

El conjunto de datos ya se encuentra dividido en subconjuntos de entrenamiento (28,709 imágenes) y prueba (7,178 imágenes), y las imágenes están organizadas jerárquicamente por clase, lo que permite un flujo de trabajo compatible con herramientas como `ImageDataGenerator` de Keras¹.

Durante el preprocessamiento, se han aplicado técnicas de redimensionado (en caso necesario) y normalización de valores, así como operaciones de aumento de datos (*data augmentation*) como rotación aleatoria, inversión horizontal y ajuste de brillo, con el objetivo de mejorar la capacidad de generalización del modelo y prevenir el sobreajuste.

En la Figura 4.8 se muestran ejemplos representativos de cada una de las clases emocionales incluidas en el dataset.

¹Este módulo permite preparar y transformar imágenes automáticamente mientras se entrena el modelo. Por ejemplo, puede girarlas, cambiarlas de tamaño o reflejarlas horizontalmente, lo que ayuda a que el modelo aprenda mejor sin necesidad de más datos.

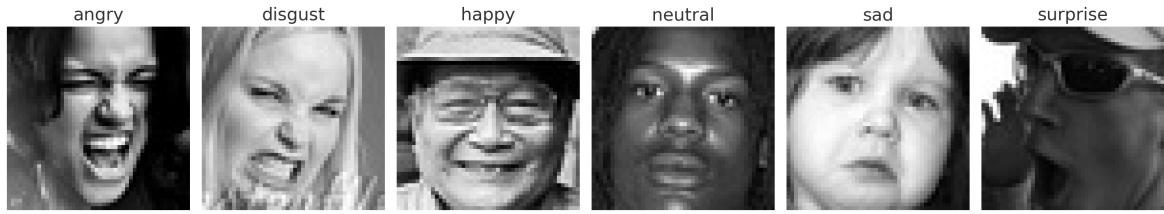


Figura 4.8: Ejemplos de imágenes por clase en el dataset de emociones FER2013.

Asimismo, se ha analizado la distribución de clases tanto para el conjunto de entrenamiento como el de prueba. Los resultados se presentan en las Figuras 4.9 y 4.10, respectivamente. Se observa un cierto grado de desbalanceo, especialmente en clases como *disgust* o *surprise*, lo que justifica el uso de técnicas de aumento de datos. La Figura 4.11 permite visualizar de forma comparativa las proporciones por clase entre ambos conjuntos antes del *data augmentation*, el conjunto de entrenamiento y el de prueba. Aunque estas técnicas de aumento de datos no modifican la distribución original almacenada, sí generan variaciones sintéticas durante el entrenamiento, lo que contribuye a reducir el sesgo y mejorar la generalización del modelo. Por tanto, no existe una nueva distribución permanente de clases, pero sí un enriquecimiento efectivo del conjunto de entrenamiento en tiempo real.

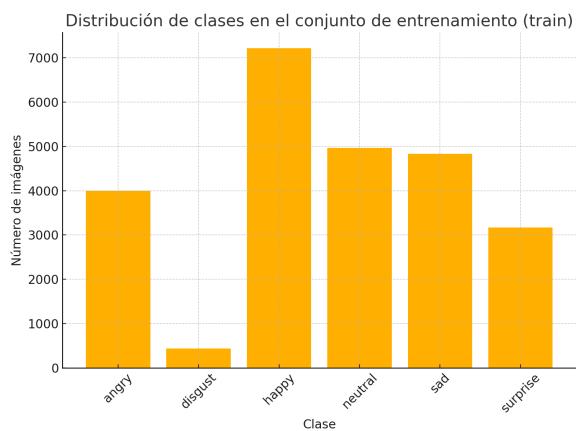


Figura 4.9: Distribución de imágenes por clase en el conjunto de entrenamiento.

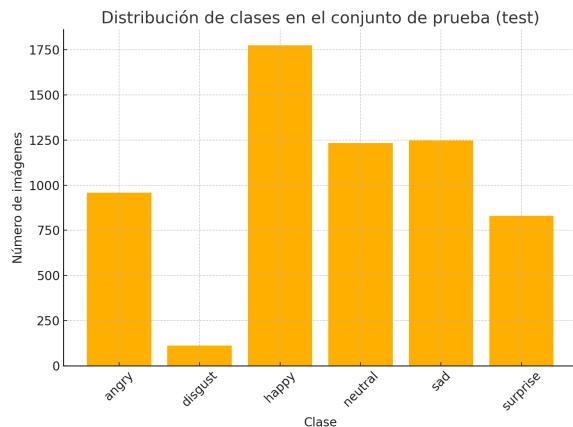


Figura 4.10: Distribución de imágenes por clase en el conjunto de prueba.

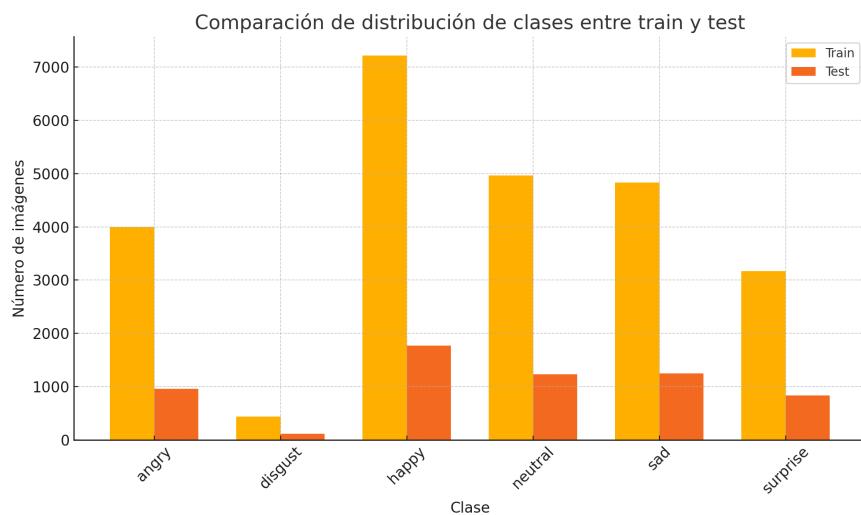


Figura 4.11: Comparativa entre las distribuciones del conjunto de entrenamiento y prueba.

4.3.3 Entrenamiento del Modelo GAN

El diseño y entrenamiento del modelo GAN se ha desarrollado a través de un conjunto de etapas sucesivas, cada una enfocada en un objetivo específico dentro del ciclo de experimentación, validación y adaptación progresiva al dominio de interés. Este enfoque iterativo ha permitido refinar tanto la arquitectura como los hiperparámetros, garantizando una convergencia estable y resultados coherentes en diferentes escenarios.

Progresión en la resolución de imágenes y arquitectura

El desarrollo del sistema basado en GANs se inició como una prueba de concepto utilizando imágenes de baja resolución (28×28 píxeles), con el objetivo de validar la arquitectura y asegurar la estabilidad del entrenamiento en un entorno controlado. Para ello, se empleó

el conjunto de datos MNIST, ampliamente utilizado en tareas de generación debido a su simplicidad y formato unicanal. Una vez verificado el correcto funcionamiento del sistema en esta configuración básica, se incrementó progresivamente la resolución de las imágenes generadas y discriminadas, con el propósito de adaptarse a un dominio más realista centrado en rostros humanos.

En esta segunda etapa se seleccionó una resolución de 52×52 píxeles, que permitió reducir significativamente el tiempo de entrenamiento por época y la ocupación de memoria en Unidad de Procesamiento Gráfico (Graphics Processing Unit) (GPU) en comparación con configuraciones más exigentes. Si bien no se registraron métricas exactas de carga computacional en esta fase, la diferencia de rendimiento fue perceptible y facilitó una experimentación más ágil durante los primeros ensayos.

Finalmente, en la configuración experimental definitiva se adoptó una resolución de 96×96 píxeles, empleando imágenes extraídas de fotogramas de vídeo correspondientes a rostros impresos. Este incremento de resolución implicó un ajuste proporcional en las dimensiones de entrada y salida del generador y discriminador, así como en los tamaños internos de las capas densas. Aunque la arquitectura se mantuvo en su forma secuencial y relativamente simple, fue necesario reevaluar ciertos hiperparámetros y adaptar el preprocessamiento de datos para garantizar la coherencia en la distribución de los valores de entrada.

Visualización del progreso. Para evaluar visualmente el comportamiento de la red a lo largo del entrenamiento, se generaron muestras sintéticas cada 100 épocas utilizando una cuadrícula de 16 imágenes (4x4). En la Figura 4.12, se muestra una secuencia representativa de estas imágenes, generadas en distintos momentos del entrenamiento (desde la época 0 hasta la 900).

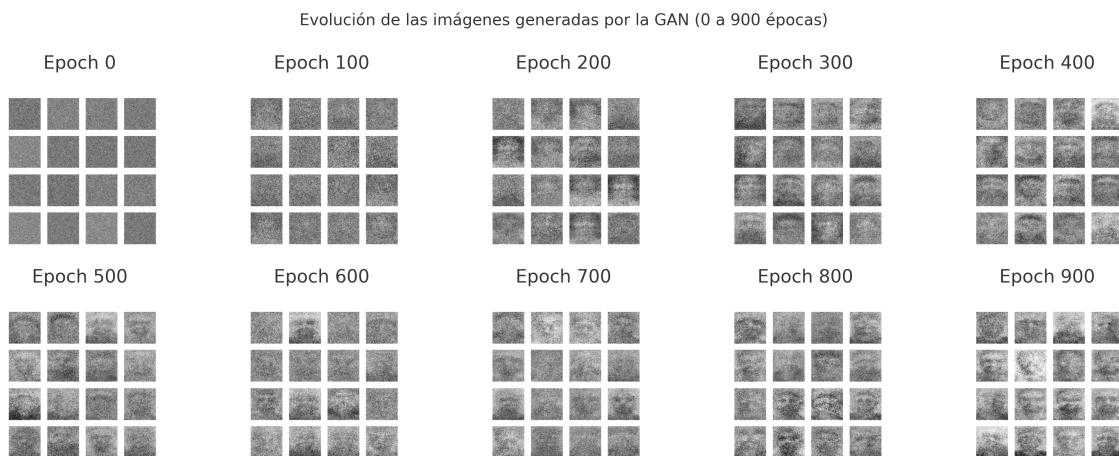


Figura 4.12: Evolución de las imágenes generadas por la GAN a lo largo del entrenamiento, desde la época 0 hasta la 900. Puede observarse una mejora progresiva en la definición y estructura de los rostros sintéticos.

Comparativa intermedia. Para complementar la evaluación visual del entrenamiento, se

ha generado una imagen comparativa con muestras correspondientes a las épocas 0, 400 y 900. En la Figura 4.13, se observa con claridad cómo la calidad y estructura de los rostros sintéticos mejora significativamente con el avance del entrenamiento.

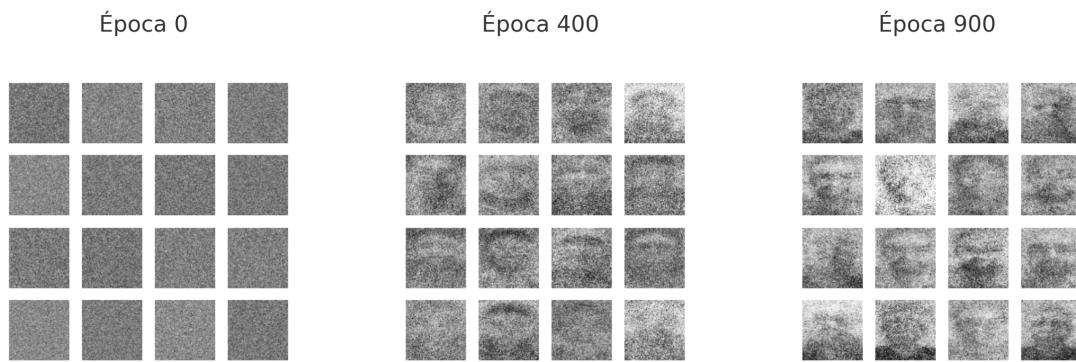


Figura 4.13: Comparativa de la calidad de las imágenes generadas por la GAN en tres momentos clave del entrenamiento (épocas 0, 400 y 900). La mejora en la coherencia visual y los rasgos faciales es evidente.

Visualización ampliada. Aunque la Figura 4.13 muestra la evolución de calidad en las imágenes generadas por la GAN en distintas épocas, las figuras compactas dificultan apreciar algunos detalles faciales relevantes. Por este motivo, a continuación se presentan de forma independiente y ampliada las muestras correspondientes a las épocas 0, 400 y 900, con el fin de observar más claramente la evolución visual del modelo.

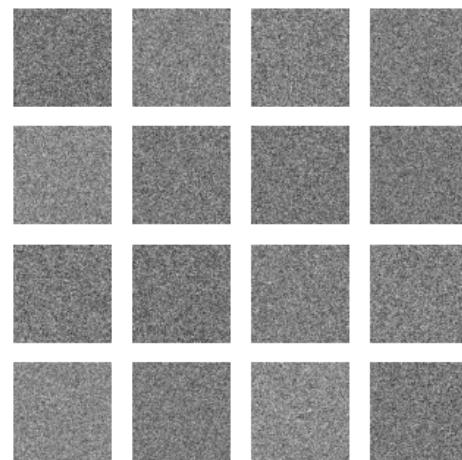


Figura 4.14: Imagen generada por la GAN en la época 0. Aún no se distinguen rasgos faciales definidos.

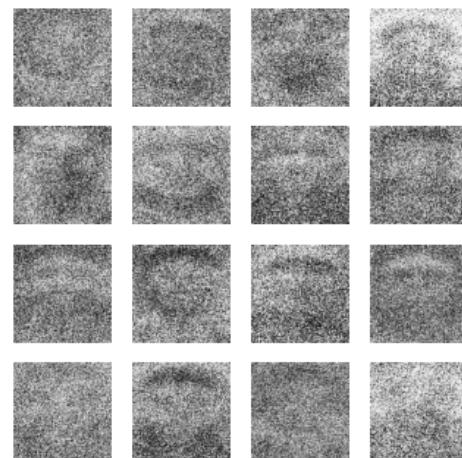


Figura 4.15: Imagen generada por la GAN en la época 400. Comienzan a apreciarse estructuras faciales.

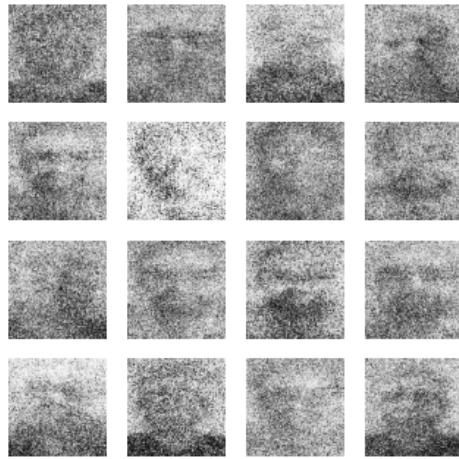


Figura 4.16: Imagen generada por la GAN en la época 900. Los rostros sintéticos presentan mayor definición y coherencia.

4.3.3.1 Etapa 1: Validación funcional con MNIST

Como punto de partida, se diseñó una arquitectura inicial orientada a validar la correcta implementación del ciclo de entrenamiento generador-discriminador. Para ello, se empleó el conjunto de datos MNIST, ampliamente utilizado en entornos de pruebas por su bajo nivel de complejidad. Las imágenes de dígitos manuscritos, en escala de grises y resolución 28×28 píxeles, permitieron evaluar si el generador era capaz de aprender patrones visuales básicos y si el discriminador podía establecer una separación efectiva entre imágenes reales y sintéticas.

El modelo, desarrollado en el script `gan.py`, utilizaba redes densas tanto para el generador como para el discriminador, con activaciones LeakyReLU, normalización por lotes y funciones de pérdida binaria cruzada. Las imágenes generadas se guardaban cada 100 épocas para facilitar una inspección cualitativa del progreso. Esta etapa resultó clave para verificar la estabilidad numérica del entrenamiento y ajustar parámetros como el tamaño del vector latente ($z \in \mathbb{R}^{100}$) y la tasa de aprendizaje del optimizador Adam.

4.3.3.2 Etapa 2: Generación de rostros desde vídeos de ataques por impresión

Una vez validado el modelo en un entorno controlado, se abordó su adaptación a un escenario más realista: la generación de imágenes de rostros humanos obtenidas a partir de vídeos de ataques por impresión. En esta etapa, el objetivo fue simular las condiciones visuales propias de ataques físicos mediante la extracción de fotogramas de vídeos con rostros impresos en papel.

El script `gan2.py` introdujo una arquitectura adaptada a imágenes de mayor resolución (96×96 píxeles), manteniendo la estructura de red densa pero ajustando la dimensión de entrada y salida para adaptarse al nuevo formato. La función `load_data_from_videos` permitió extraer los 100 primeros fotogramas de cada archivo de vídeo y preprocesarlos en escala

de grises, alineando así el formato de entrada con las necesidades del modelo.

Durante el entrenamiento, se registraron las métricas de pérdida tanto del generador como del discriminador, y se almacenaron los pesos correspondientes a los mejores valores observados, lo que permitió preservar configuraciones óptimas de cada red. Asimismo, se generaron y guardaron muestras visuales en la carpeta `generated_images` para evaluar la evolución de las imágenes generadas en distintos puntos del proceso.

4.3.3.3 Etapa 3: Generación sobre expresiones faciales (FER2013)

Con el objetivo de aplicar la GAN a un dominio directamente relacionado con la clasificación de emociones a partir de expresiones faciales, se entrenó el modelo sobre el conjunto de datos FER2013. Este dataset contiene imágenes de rostros etiquetados con emociones, todas en escala de grises y con resolución 48×48 píxeles, lo que lo convierte en una base adecuada para validar la capacidad del generador para sintetizar expresiones faciales plausibles.

El script `gan3.py` incluyó un módulo de carga de datos a partir de carpetas organizadas por clase emocional, permitiendo procesar el conjunto de entrenamiento completo. Las imágenes eran preprocesadas mediante normalización al rango $[-1, 1]$, y el entrenamiento se desarrolló siguiendo la misma lógica que en etapas anteriores: alternancia entre generador (G) y discriminador (D), generación periódica de muestras, y almacenamiento de pesos óptimos.

Esta etapa representó un punto intermedio clave, al combinar imágenes con mayor complejidad estructural (rostros) y variabilidad semántica (emociones), sirviendo de puente hacia una aplicación práctica del modelo en el contexto del sistema de detección.

4.3.3.4 Etapa 4: Evaluación sistemática y control de sobreajuste

Finalmente, se desarrolló una versión avanzada del proceso de entrenamiento (`gan4.py`) que incorporaba mecanismos adicionales de validación sobre un conjunto de test. Esta estrategia tenía como propósito evaluar la capacidad del discriminador para generalizar más allá de los ejemplos vistos durante el entrenamiento, así como verificar que el generador no incurría en problemas de colapso de modo o sobreajuste visual.

Cada 100 épocas se ejecutaba un módulo de evaluación del discriminador mediante una muestra aleatoria del conjunto de test, y se registraba su precisión media al clasificar tanto imágenes reales como generadas. Además, se continuó el almacenamiento de imágenes generadas a lo largo del entrenamiento, posibilitando una evaluación cualitativa continua de la calidad de síntesis.

Este sistema de evaluación cruzada permitió establecer un control efectivo sobre la evolución del aprendizaje, aportando trazabilidad y evidencia empírica del rendimiento del modelo.

4.3.4 Estrategia de Evaluación

Para evaluar el comportamiento del sistema propuesto, se ha planteado una estrategia de validación basada en múltiples ejecuciones sobre distintos datasets y resoluciones. En particular, se han analizado:

- La estabilidad de las curvas de pérdida y precisión durante el entrenamiento.
 - La calidad visual de las imágenes generadas.
-

- La capacidad de generalización del discriminador frente a ejemplos no vistos (evaluación cruzada (*cross-validation*)).

Los resultados de esta evaluación se presentan en detalle en el Capítulo 5, donde se analizan comparativamente las salidas de los scripts `gan2.py`, `gan3.py` y `gan4.py`.

4.4 Sistema de Detección Basado en Profundidad

Con el objetivo de reforzar la detección de intentos de suplantación facial mediante imágenes impresas, se ha desarrollado un sistema alternativo basado en el análisis estructural de las imágenes, empleando mapas de profundidad generados a partir de una única imagen RGB.

Para esta tarea se ha utilizado el modelo **MiDaS** (*Mixed Depth Architecture Search*) Ranftl y cols. (2020), una red neuronal profunda desarrollada por Intel que permite estimar mapas de profundidad relativa a partir de imágenes estáticas. Este modelo ha sido preentrenado sobre un conjunto amplio y heterogéneo de datasets, lo que le confiere una notable capacidad de generalización a contextos visuales variados, sin necesidad de calibración ni datos adicionales de escena.

A diferencia de los sensores físicos de profundidad, MiDaS genera una estimación monocular —no métrica— que refleja relaciones espaciales relativas entre objetos de la escena. Esta propiedad resulta especialmente útil para distinguir imágenes planas (como las fotocopias) de imágenes con volumen (rostros reales), ya que en estas últimas suelen observarse mayores variaciones de profundidad y complejidad estructural.

El sistema desarrollado sigue un flujo dividido en tres fases principales:

- 1. Estimación del mapa de profundidad:** A partir de una imagen RGB, se aplica el modelo MiDaS (versión `dpt_large.pt`) para generar un mapa de profundidad relativo. Este mapa se normaliza al rango $[0, 255]^2$ y se visualiza con un colormap tipo JET³, facilitando así su interpretación visual.
- 2. Extracción de región facial (ROI):** Sobre la imagen original o su correspondiente mapa de profundidad se localiza el rostro mediante dos estrategias complementarias: detección automática con Haar Cascade⁴ y segmentación morfológica por contornos⁵. Se obtiene así una máscara binaria que permite aislar la región de interés sobre la cual se realizarán las mediciones.
- 3. Análisis de textura y profundidad:** Dentro del ROI extraído se calcula la varianza del Laplaciano⁶ (como indicador de textura), así como métricas estadísticas relacionadas

²La normalización al rango $[0, 255]$ convierte los valores flotantes de profundidad en una escala estándar de 8 bits, adecuada para su visualización como imagen en escala de grises o color.

³El colormap JET asigna colores desde azul (valores bajos) hasta rojo (valores altos), lo que facilita identificar visualmente las diferencias de profundidad en una imagen.

⁴Haar Cascade es un método de detección de objetos basado en clasificadores entrenados con características Haar, comúnmente utilizado en OpenCV para detectar rostros.

⁵La segmentación morfológica por contornos aplica operaciones sobre una imagen binaria para identificar los bordes de objetos, como la silueta del rostro.

⁶La varianza del Laplaciano es una métrica común para estimar la nitidez o nivel de textura de una imagen, siendo útil para detectar desenfoque o superficies planas.

con la profundidad (mínimo, máximo, rango y desviación típica)⁷. Estas características permiten comparar imágenes reales con sus versiones falsificadas y detectar posibles patrones indicativos de planitud.

La comparación entre imágenes reales y falsificadas se ha realizado por pares. En cada caso, se han procesado las imágenes originales y sus correspondientes photocopies para obtener mapas de profundidad y regiones faciales. Posteriormente, se han analizado cuantitativamente las diferencias en variabilidad estructural y complejidad superficial. Como apoyo a este análisis, se han generado histogramas de distribución de profundidad en la región de la cara, lo que ha facilitado la identificación de características distintivas en las copias impresas.

Este enfoque se ha implementado mediante una colección de scripts en Python que automatizan el procesamiento por lotes, la estimación de profundidad, la extracción de máscaras y la visualización de resultados. Gracias a ello, el sistema puede evaluar nuevas imágenes de forma reproducible y con un costo computacional moderado.

En conjunto, esta línea metodológica ha demostrado un gran potencial para complementar sistemas clásicos de detección basados únicamente en información RGB, aportando una capa estructural adicional que puede resultar clave para la detección de ataques por presentación.

4.4.1 Procesamiento de Imágenes con MiDaS

Con el objetivo de estimar mapas de profundidad a partir de imágenes faciales RGB, se ha empleado el modelo **MiDaS** (Monocular Depth Estimation) en su versión DPT-Large, previamente entrenado sobre grandes conjuntos de datos. MiDaS cuenta actualmente con varias versiones, entre las que destacan DPT-Hybrid, DPT-Large y MiDaS v2.1 Small, que difieren principalmente en términos de precisión y eficiencia computacional. En este trabajo se ha optado por DPT-Large por su mayor resolución espacial y calidad en la estimación de profundidad relativa, características especialmente relevantes en contextos biométricos donde los detalles faciales finos son esenciales para detectar diferencias estructurales.

Este modelo permite obtener una representación relativa de la geometría tridimensional de una escena a partir de una única imagen, sin necesidad de sensores adicionales.

El proceso de estimación ha sido automatizado mediante un script en Python denominado `run_midas.py`, diseñado para ejecutarse en un entorno con soporte GPU en el servidor remoto. El flujo de trabajo seguido por este script se resume de forma clara y visual en la Tabla 4.17.

⁷Estas métricas permiten cuantificar la variación espacial en la profundidad, lo que puede indicar si una superficie tiene volumen o es plana.

1. Cargar imagen	Se abre la imagen que se desea analizar (por ejemplo, una foto de un rostro).
2. Preparar la imagen	<ul style="list-style-type: none"> • Se convierte al formato RGB (si es necesario). • Se redimensiona y adapta al formato que requiere MiDaS (“transformación”).
3. Aplicar el modelo	MiDaS analiza la imagen y genera un mapa de profundidad.
4. Visualizar el resultado	<ul style="list-style-type: none"> • Se normalizan los valores de profundidad al rango 0–255. • Se puede aplicar una paleta de colores (COLORMAP_JET) para facilitar su visualización.
5. Guardar el resultado	La imagen con el mapa de profundidad coloreado se guarda automáticamente.

Figura 4.17: Flujo simplificado del procesamiento de una imagen con MiDaS.

A modo de ejemplo visual, la Figura 4.18 muestra una imagen original y su correspondiente mapa de profundidad generado con MiDaS.



Figura 4.18: Imagen de entrada (izquierda) y su respectivo mapa de profundidad coloreado (derecha) generado con MiDaS.

La ejecución del script es interactiva: solicita al usuario el nombre de la imagen a procesar, realiza la inferencia y almacena automáticamente el resultado sin intervención adicional.

Además del script principal, se implementaron dos variantes adicionales con fines de prueba:

- `run_midas2.py`: versión que permite forzar la conversión del modelo a precisión Punto

Flotante de 32 Bits (32-bit Floating Point) (FP32)⁸ y definir la imagen a procesar desde el propio script.

- `run_midas-color.py`: variante diseñada para generar directamente mapas de profundidad con codificación en color Jet⁹.

Este procesamiento constituye la base para los análisis de comparación aplicados posteriormente sobre imágenes reales y sus fotocopias. Todos los resultados fueron almacenados en formato .jpg y transferidos a la máquina local mediante comandos Protocolo Seguro de Copia (Secure Copy Protocol) (SCP).

4.4.2 Extracción de ROI y Análisis de Textura

Tras la generación de los mapas de profundidad con MiDaS, el siguiente paso ha consistido en aislar la ROI correspondiente al rostro humano y analizar sus propiedades geométricas y texturales. Este proceso permite obtener métricas cuantitativas que diferencian imágenes reales de sus posibles fotocopias, contribuyendo así a la detección de intentos de suplantación.

Para la extracción de la ROI se han utilizado dos enfoques complementarios:

- **Detección por contornos**: mediante umbralización binaria y extracción del mayor contorno externo, asumido como la silueta de la cabeza.
- **Detección facial automática**: basada en clasificadores Haar Cascade de OpenCV, aplicados sobre la imagen original en color para generar una máscara localizada del rostro.

Estas máscaras se han aplicado sobre las imágenes de profundidad normalizadas para recortar exclusivamente la información correspondiente al área facial. La Figura 4.19 ilustra un ejemplo del proceso de detección y extracción de la ROI a partir de una máscara facial.

⁸FP32 (punto flotante de 32 bits) es una representación estándar utilizada en el entrenamiento e inferencia de modelos de redes neuronales. Ofrece un equilibrio entre precisión numérica y velocidad computacional.

⁹El colormap Jet representa los valores numéricos usando una escala de colores que va de azul (valores bajos) a rojo (valores altos), facilitando la interpretación visual de los mapas de profundidad.

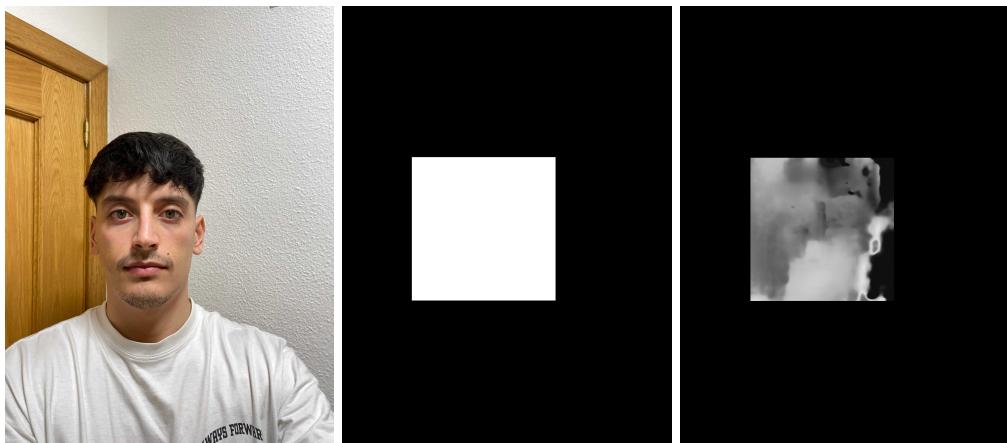


Figura 4.19: Ejemplo de extracción de ROI: imagen original (izquierda), máscara facial generada (centro) y ROI aplicada sobre el mapa de profundidad (derecha).

Una vez extraída la región de interés, se han calculado tres métricas fundamentales sobre la información de profundidad:

- **Varianza del Laplaciano:** como medida de textura, asociada a la nitidez local del mapa.
- **Desviación típica de la profundidad:** indica la dispersión de valores dentro de la ROI.
- **Rango de profundidad:** diferencia entre los valores máximo y mínimo de la región analizada.

Estos cálculos se han realizado mediante los scripts `analyze.py` e `imageDepthContourAnalyzer.py`, ambos diseñados para ejecutarse en el servidor con entorno GPU. La ejecución permite comparar dos imágenes —una real y su posible fotocopia— y evaluar la autenticidad de esta última con base en umbrales establecidos empíricamente (15 para la variación de profundidad y 50 para la textura).

Para visualizar los resultados obtenidos, se ha aplicado el proceso descrito previamente —detección de ROI y extracción de métricas— sobre pares de imágenes compuestos por una foto original y su correspondiente fotocopia.

Los valores de profundidad dentro de cada ROI se han representado en forma de histogramas comparativos, que permiten apreciar las diferencias estructurales entre ambas imágenes. Estos histogramas se han almacenado automáticamente en la carpeta `imagenes/graphs`. Asimismo, las máscaras faciales generadas y las regiones recortadas del rostro (ROI) se han guardado en la carpeta `imagenes/roi` para su posterior inspección visual.

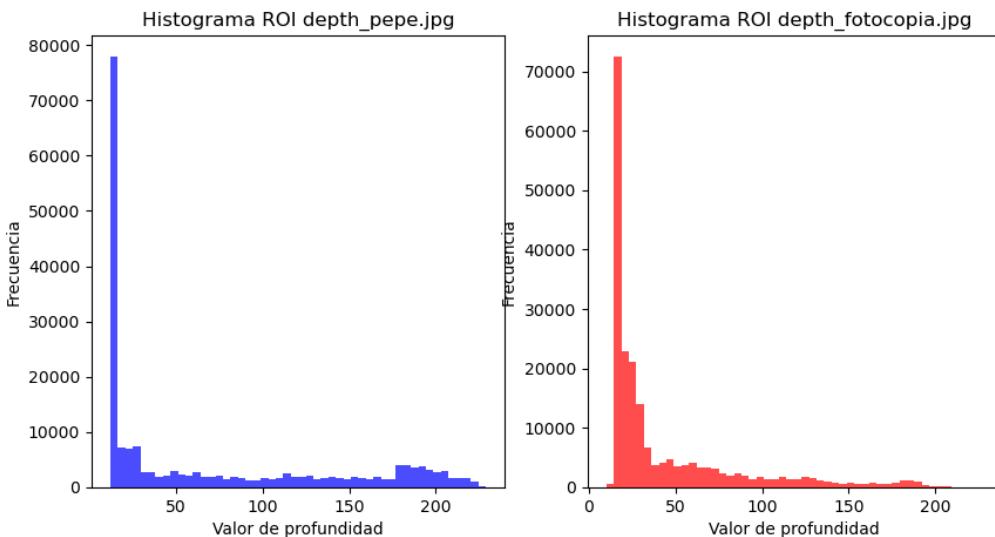


Figura 4.20: Histogramas de profundidad de las regiones de interés. Izquierda: imagen real. Derecha: fotocopia.

Este análisis de textura y profundidad ha demostrado ser eficaz en múltiples casos para detectar diferencias sutiles entre una imagen original y una copia impresa. En particular, la fotocopia suele presentar una menor variabilidad en profundidad y una textura suavizada, debido a las limitaciones del medio físico de reproducción.

Todos los resultados han sido almacenados y registrados automáticamente, y se utilizarán en el capítulo siguiente para elaborar las conclusiones y establecer criterios de detección basados en umbrales medidos empíricamente.

4.4.3 Comparación entre Imagen Real y Fotocopia

Para avanzar en la detección de posibles suplantaciones mediante el uso de fotocopias, se ha diseñado y ejecutado un procedimiento de comparación entre imágenes reales y sus correspondientes reproducciones impresas. Este análisis se ha basado en la aplicación del modelo MiDaS para la estimación monocular de profundidad, seguido de un estudio cuantitativo de la región facial, extraída a partir de técnicas de segmentación automática.

Procesamiento y Enfoques de Extracción de ROI

Cada imagen ha sido procesada de manera independiente, generando un mapa de profundidad normalizado. A partir de dicho mapa, se ha delimitado una región de interés (ROI) centrada en el rostro, mediante dos estrategias complementarias:

- **Segmentación por contornos:** este método ha partido de una binarización del mapa de profundidad, sobre la cual se han detectado los contornos externos. Se ha asumido que el contorno más grande corresponde a la silueta de la cabeza, permitiendo así extraer de forma robusta una máscara aproximada de la región craneofacial. Este enfoque resulta

especialmente útil en imágenes donde el rostro no ha sido claramente captado por detectores automáticos.

- **Detección facial con Haar Cascades:** se ha aplicado un clasificador Haar preentrenado sobre la imagen original en color. Este método ha permitido localizar de forma precisa la zona facial mediante coordenadas rectangulares, generando una máscara que posteriormente se ha trasladado al mapa de profundidad. Esta técnica, basada en aprendizaje automático, ofrece mayor precisión en contextos controlados, aunque puede fallar en imágenes desenfocadas o con expresiones atípicas.

Ambas estrategias han sido implementadas en scripts específicos y se han complementando entre sí, permitiendo comparar los resultados obtenidos desde distintas perspectivas de segmentación.

Métricas Analizadas

Sobre las regiones de interés extraídas se han calculado diferentes métricas, destinadas a evaluar tanto la geometría como la riqueza estructural de la superficie facial:

- **Variación de profundidad:** se ha definido como la diferencia entre los valores máximo y mínimo del mapa dentro de la ROI. Una mayor variabilidad sugiere una estructura tridimensional rica, típica de un rostro real. Las fotocopias, al ser superficies planas, tienden a presentar valores más uniformes.
- **Textura (Laplaciano):** se ha calculado la desviación típica del Laplaciano de la ROI. Esta métrica captura la cantidad de detalle local (bordes, rugosidades), y su disminución es un indicador de superficies suavizadas, como ocurre en impresiones en papel.
- **Desviación estándar de la profundidad:** proporciona una medida global de dispersión de los valores dentro de la ROI, complementando la variación y permitiendo detectar irregularidades no visibles a simple vista.
- **Distribución de valores:** se han generado histogramas de los niveles de profundidad presentes en cada ROI, que permiten una visualización directa de la dispersión, los picos y los valores dominantes. Estas gráficas han sido clave para identificar patrones repetitivos en imágenes impresas.

Uso de Umbrales y Consideraciones

Durante la fase de análisis, se han utilizado umbrales de referencia orientativos, no como límites de decisión estrictos, sino como herramientas exploratorias para resaltar diferencias significativas entre pares de imágenes. Esta aproximación ha permitido observar patrones relevantes sin asumir reglas fijas, lo cual es esencial dada la variabilidad inherente a las condiciones de captura y a las características individuales de cada sujeto.

Resultados y Visualización

El procedimiento completo ha sido automatizado mediante scripts como `analyze.py`, `imageDepthAnalyzer.py` e `imageDepthContourAnalyzer.py`, los cuales permiten procesar imágenes, extraer métricas y generar gráficas comparativas. Los resultados han sido organizados sistemáticamente en las carpetas `imagenes/output`, `imagenes/roi` y `imagenes/graphs`.

La Figura 4.21 muestra un ejemplo de comparación entre los histogramas de profundidad de una imagen real y su fotocopia. Como puede observarse, la imagen impresa tiende a concentrar los valores en un rango más reducido, con menos dispersión y textura.

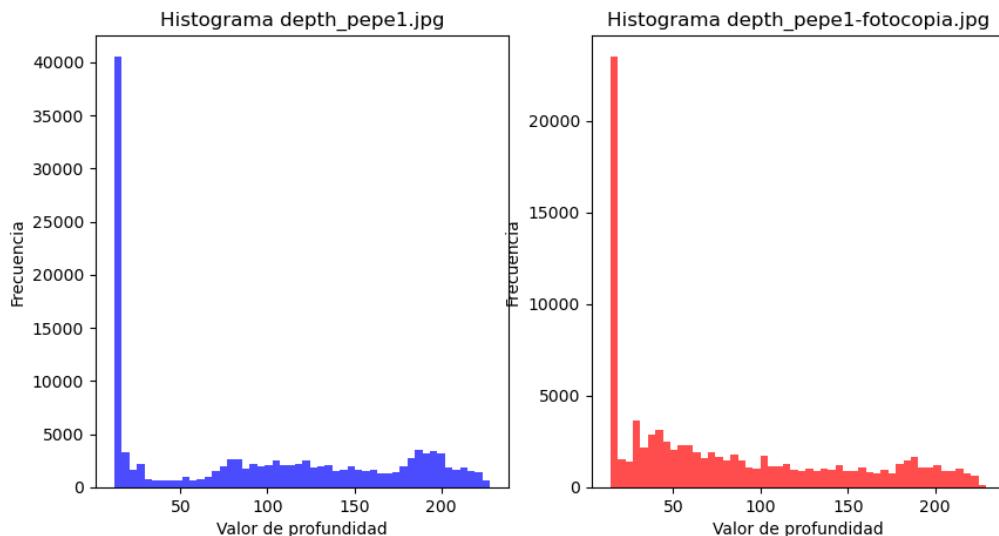


Figura 4.21: Histogramas de profundidad en la región facial. Izquierda: imagen original. Derecha: fotocopia.

Adicionalmente, en la Figura 4.23 se presenta una comparación visual de las ROIs extraídas con ambos métodos (contorno vs. detección facial), sobre un mismo par de imágenes. Esta comparación permite apreciar las diferencias prácticas entre ambos enfoques de segmentación.

Antes de mostrar los análisis cuantitativos, en la Figura 4.22 se presenta la imagen original y su correspondiente versión fotocopiada, utilizadas en las comparaciones siguientes. Esto permite visualizar el punto de partida del análisis.



Figura 4.22: Imagen original (izquierda) y su versión fotocopiada (derecha) empleadas en el análisis de profundidad.

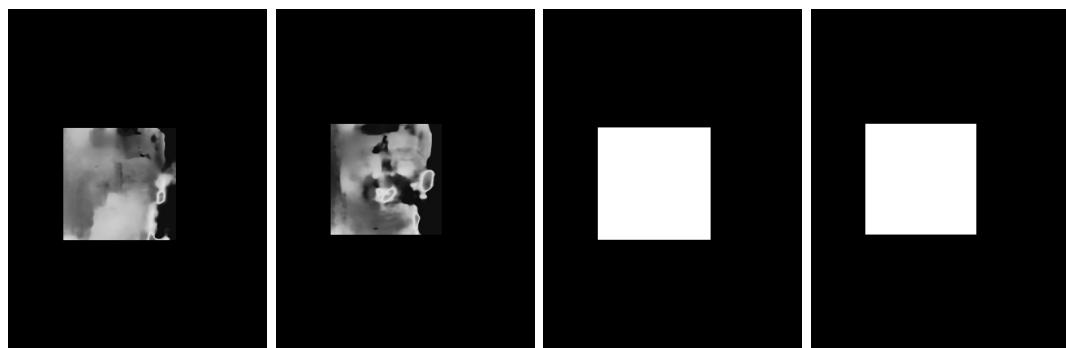


Figura 4.23: Comparativa de ROIs: recortes de profundidad (izquierda) y máscaras faciales Haar (derecha), para imagen real y fotocopia.

Aunque no se ha establecido una regla de decisión definitiva, los patrones observados sugieren que las imágenes de fotocopias presentan, con frecuencia, menor complejidad estructural en sus mapas de profundidad. La reducción de textura y la homogeneidad en la distribución de niveles han sido indicadores frecuentes de reproducción impresa. Estos indicios, combinados con futuras técnicas de aprendizaje automático, podrían convertirse en la base de un sistema automático de detección de fraudes visuales.

4.5 Herramientas y Entorno de Desarrollo

Esta sección describe los recursos técnicos y la estructura del entorno empleada a lo largo del desarrollo del Trabajo Fin de Máster. Se detallan las herramientas utilizadas en las distintas fases experimentales, incluyendo el entrenamiento de redes GAN, el análisis de profundidad monocular, la gestión del entorno de ejecución, así como la organización y automatización del proyecto.

4.5.1 Librerías y Frameworks

El desarrollo ha requerido el uso combinado de múltiples librerías y frameworks especializados en visión por computador, aprendizaje profundo y procesamiento de datos:

- **TensorFlow** y **Keras**: utilizados en la fase inicial para implementar redes GAN básicas.
- **PyTorch**: adoptado como framework principal para etapas posteriores por su flexibilidad y compatibilidad con modelos avanzados como MiDaS.
- **NumPy**, **Matplotlib** y **OpenCV**: fundamentales para la manipulación de matrices, visualización de resultados y análisis de imágenes.
- **MiDaS**: modelo de código abierto basado en vision transformers, empleado para la generación de mapas de profundidad monocular.

4.5.2 Entorno de Ejecución

El proyecto ha sido desarrollado y ejecutado en un servidor remoto con GPU, al que se ha accedido mediante terminal vía **Protocolo Seguro de Conexión Remota** (**Secure Shell**) (**SSH**). La gestión de las dependencias se ha realizado mediante un entorno virtual de **conda**, llamado **myenv**, que ha permitido encapsular las versiones específicas de librerías necesarias.

Un ejemplo representativo del entorno es el siguiente:

```
(myenv) joseburgos@servergpu:~/tfm$ ls  
fase2  fase3  fase4  fase5  fase6  __pycache__  spec-file.txt
```

4.5.3 Estructura del Proyecto

El proyecto se ha organizado siguiendo una estructura jerárquica basada en fases, donde cada carpeta contiene scripts, modelos y resultados específicos de cada etapa del desarrollo experimental. A continuación, se presenta un esquema representativo de dicha organización:

tfm/

```
|-- fase2/
|   |-- gan.py
|
|-- fase3/
|   |-- gan2.py
|   |-- dataset/
|       |-- print_samples/
|           |-- print_1.mp4
|           |-- ...
|   |-- generated_images/
|       |-- epoch_0.png
|       |-- epoch_100.png
|
|-- fase5/
|   |-- gan3.py
|   |-- gan4.py
|   |-- dataset/
|       |-- train/
|           |-- happy/
|               |-- Training_12345.jpg
|   |-- generated_images/
|       |-- epoch_0.png
|       |-- loss_plot_epoch_100.png
|
|-- fase6/
|   |-- MiDaS/
|   |-- imagenes/
|       |-- input/
|       |-- output/
|       |-- roi/
|       |-- graphs/
|   |-- scripts/
|       |-- run_midas.py
|       |-- analyze.py
|
|-- spec-file.txt
|-- install_dependencies.sh
```

Cada fase incluye sus propios scripts, modelos y datasets asociados, manteniéndose independientes en cuanto a dependencias y datos intermedios.

4.5.4 Intercambio de Archivos

Se ha realizado una transferencia frecuente de resultados y archivos entre el entorno local y el remoto, utilizando el protocolo SCP. Este intercambio ha permitido visualizar localmente imágenes generadas, pesos entrenados o histogramas obtenidos en las fases de evaluación.

4.5.5 Reproducibilidad

Para asegurar la reproducibilidad del entorno de trabajo, se ha creado un archivo de configuración de entorno conda (`environment.yaml`) y un script de instalación de dependencias para la fase de profundidad (`install_dependencies.sh`). Estos archivos permiten replicar el entorno en otras máquinas compatibles.

4.5.6 Visualización y Guardado de Resultados

Durante las fases de entrenamiento de las GANs, se han generado imágenes sintéticas y gráficos de evolución de pérdida (*loss plots*) en intervalos regulares. Igualmente, en las fases de análisis de profundidad, se han almacenado mapas de profundidad codificados en color, máscaras faciales, regiones de interés (ROI) y sus histogramas asociados.

Ejemplo de carpetas generadas automáticamente:

```
fase5/generated_images/
  epoch_0.png
  epoch_100.png
  loss_plot_epoch_100.png
  ...

fase6/imagenes/
  input/
  output/
  roi/
  graphs/
```

4.5.7 Automatización y Scripts Personalizados

Para gestionar la carga de datos, entrenar modelos, generar salidas visuales y realizar comparaciones automatizadas entre imágenes reales y fotocopiadas, se han desarrollado distintos scripts personalizados, entre los que destacan:

- `gan.py`, `gan2.py`, `gan3.py`, `gan4.py`: versiones evolutivas del sistema GAN.
 - `run_midas.py`: inferencia de profundidad con MiDaS.
 - `analyze.py` e `imageDepthAnalyzer.py`: scripts de análisis comparativo entre imágenes reales y falsificadas.
-

Cada script ha sido documentado mediante comentarios en el código y estructurado para facilitar su ejecución modular.

4.6 Limitaciones y Desafíos Encontrados

A lo largo del desarrollo del trabajo, se han identificado diversos obstáculos que han condicionado el diseño experimental, el análisis de resultados y la estabilidad de los sistemas implementados. Esta sección recoge de forma estructurada las principales limitaciones y problemas detectados, organizados según los ejes fundamentales del proyecto.

4.6.1 Dificultades en la Línea GAN

Durante el desarrollo de los modelos generativos basados en redes antagónicas (GANs), se han presentado diversas dificultades técnicas que han condicionado la estabilidad y eficacia del entrenamiento. Una de las más destacadas ha sido la aparición recurrente del fenómeno conocido como *colapso de modo* (*mode collapse*). Este problema ocurre cuando el generador aprende a producir un conjunto muy limitado de ejemplos que, aunque logran engañar al discriminador, pierden diversidad. En la práctica, esto se ha traducido en secuencias de entrenamiento donde el generador produce imágenes muy similares entre sí, impidiendo capturar adecuadamente la variabilidad presente en los datos reales.

Otro desafío importante ha sido lograr un entrenamiento estable y progresivo. Las GANs son notoriamente sensibles a pequeños ajustes en los hiperparámetros, las arquitecturas de red y las condiciones iniciales. A lo largo del trabajo se ha tenido que reajustar repetidamente la configuración de los modelos, incluyendo aspectos como la tasa de aprendizaje, el tamaño del espacio latente, y el número de capas y unidades de cada bloque. Este proceso ha requerido una combinación de validación empírica, análisis de resultados parciales y comprensión del comportamiento de cada arquitectura.

Asimismo, se ha detectado que algunos vídeos utilizados como fuente para extraer imágenes de entrenamiento no proporcionaban una cantidad suficiente de fotogramas útiles. Esto ha limitado la capacidad del generador para aprender representaciones completas y ha obligado a realizar una preselección manual de los recursos más adecuados, lo cual ha incrementado la carga de trabajo y ha afectado la escalabilidad del proceso.

4.6.2 Limitaciones del Análisis con MiDaS

En la fase de análisis basada en mapas de profundidad mediante el uso de MiDaS, también han surgido dificultades relevantes. Una de las más significativas ha sido la detección deficiente de rostros en imágenes con baja calidad visual, ya sea por condiciones de iluminación inadecuadas, desenfoque o encuadres incorrectos. La calidad de la detección del rostro condiciona directamente la precisión en la extracción de la región de interés (ROI), y por tanto afecta negativamente a la posterior comparación de mapas de profundidad.

Por otra parte, en ciertos casos, los mapas generados por MiDaS han resultado ser excesivamente planos, especialmente en imágenes que ya de por sí presentaban poca información tridimensional. Este efecto ha dificultado la extracción de métricas significativas en esas regiones, reduciendo la fiabilidad de las comparaciones.

Adicionalmente, se ha identificado la necesidad de ajustar dinámicamente los umbrales de decisión utilizados durante el análisis. En un primer momento se establecieron ciertos valores de referencia para facilitar la interpretación de los resultados. Sin embargo, la variabilidad entre individuos, expresiones, y condiciones de captura hacen que estos valores no sean universalmente válidos. La falta de un criterio generalizable ha obligado a replantear este enfoque como una guía exploratoria más que como una norma estricta de clasificación.

4.6.3 Restricciones Técnicas y del Entorno de Ejecución

En cuanto a los recursos computacionales y el entorno de ejecución, se han experimentando varias limitaciones relevantes. En particular, se han presentado incompatibilidades entre versiones de TensorFlow y cuDNN al intentar ejecutar los modelos en GPU, incluso dentro del servidor remoto principal utilizado para el entrenamiento. Estas incompatibilidades han obligado a realizar un ajuste fino de las versiones de las librerías instaladas, lo cual ha sido un proceso laborioso que ha requerido comprobar documentación, gestionar entornos virtuales y validar la estabilidad de cada configuración.

Debido a la sensibilidad del entorno a estos cambios, se ha optado por centralizar todo el desarrollo en un único entorno remoto, utilizando herramientas como `conda` para el aislamiento de dependencias, y `SCP` para la transferencia de archivos entre sistemas. A pesar de esta decisión, la gestión de versiones ha seguido siendo un desafío continuo, especialmente cuando se combinaban distintos frameworks (como TensorFlow y PyTorch) o bibliotecas que dependían de compilaciones nativas.

Estos inconvenientes han puesto de manifiesto la importancia de la reproducibilidad y la necesidad de establecer entornos de desarrollo bien definidos, así como documentar cada paso del proceso de instalación y configuración para facilitar futuras iteraciones del trabajo o su posible extensión.

5 Resultados

Este capítulo recoge y analiza los resultados obtenidos a lo largo del proceso experimental, estructurados conforme a las dos grandes líneas de trabajo definidas en la metodología: el desarrollo y evaluación de redes generativas antagónicas (GANs), y el análisis de profundidad monocular aplicado a la detección de ataques por impresión. Ambas líneas han seguido enfoques experimentales diferenciados, pero complementarios, con el objetivo común de explorar métodos innovadores para la detección de fraudes visuales en entornos de verificación facial.

La presentación de los resultados se ha organizado en bloques claramente delimitados, permitiendo una mejor interpretación de los avances logrados en cada componente del sistema. Por un lado, se analiza el desempeño de las GANs a través de diferentes fases de entrenamiento, configuraciones y conjuntos de datos, evaluando tanto la calidad visual de las imágenes generadas como la evolución de métricas internas como la pérdida y la precisión del discriminador. Por otro lado, se exponen los hallazgos obtenidos mediante el análisis de mapas de profundidad, haciendo uso de herramientas de visualización, histogramas y métricas estructurales que permiten comparar imágenes reales con sus respectivas reproducciones impresas.

Se ha adoptado una estrategia de evaluación mixta, combinando métodos cuantitativos y cualitativos. Las métricas numéricas han permitido seguir el comportamiento de los modelos de forma objetiva y reproducible, mientras que el análisis cualitativo ha aportado una perspectiva interpretativa sobre los resultados visuales, esencial para valorar el éxito del sistema en contextos donde la variabilidad de las muestras es elevada.

Las pruebas experimentales han sido organizadas por etapas, reflejando la progresión metodológica del proyecto. Cada etapa responde a un conjunto específico de objetivos, configuraciones y criterios de evaluación, y sus resultados se han documentado de manera sistemática para facilitar la comparación entre escenarios y la extracción de conclusiones relevantes.

En conjunto, este capítulo busca ofrecer una visión detallada, crítica y estructurada de los principales logros obtenidos, así como de las limitaciones encontradas durante la fase experimental. Los análisis realizados sientan las bases para futuras mejoras del sistema y abren la puerta a nuevas líneas de investigación basadas en los hallazgos aquí presentados.

5.1 Evaluación del Sistema GAN

La evaluación del sistema basado en redes generativas antagónicas (GAN) se ha llevado a cabo mediante un enfoque mixto que combina el análisis cuantitativo de métricas internas (como la pérdida del generador y del discriminador, y/o la precisión de clasificación del discriminador) con una valoración cualitativa de las imágenes generadas en diferentes momentos del entrenamiento. Esta doble perspectiva ha permitido obtener una comprensión más rica y completa del comportamiento de los modelos entrenados.

El estudio se ha estructurado en distintas etapas experimentales, que corresponden a configuraciones específicas de entrenamiento y distintos objetivos analíticos. Esta división permite aislar los efectos de ciertas variables —como la resolución de entrada o el tipo de conjunto de datos utilizado— y facilitar la comparación entre ejecuciones. Cada etapa ha sido desarrollada mediante scripts específicos del proyecto, lo que asegura una trazabilidad clara de los resultados y favorece la reproducibilidad del trabajo.

En concreto, los scripts utilizados han sido:

- `gan2.py`: orientado al análisis de la resolución de entrada y su efecto sobre la calidad y diversidad de las imágenes generadas.
- `gan3.py`: diseñado para entrenar modelos GAN sobre el conjunto de datos FER2013, centrado en el reconocimiento de emociones basado en expresiones faciales.
- `gan4.py`: encargado de realizar evaluaciones cruzadas mediante conjuntos de test separados, con el objetivo de analizar la capacidad de generalización del discriminador.

Cada uno de estos scripts ha sido acompañado de procedimientos de generación visual periódica (cada 100 épocas), así como de almacenamiento y análisis de curvas de pérdida y precisión. A lo largo de las siguientes subsecciones se presentan los resultados obtenidos en cada etapa, junto con una discusión crítica de los hallazgos más relevantes y las dificultades encontradas durante el proceso de entrenamiento.

5.1.1 Etapa 2: Análisis por resolución sobre ataques por impresión (`gan2.py`)

Con el objetivo de evaluar el impacto de la resolución de entrada sobre la calidad visual de las muestras sintéticas y la estabilidad del entrenamiento, se ha realizado un experimento comparativo empleando tres tamaños de imagen: 28×28 , 52×52 y 96×96 píxeles. Para cada resolución se ha ejecutado el modelo de forma independiente, manteniendo constante la arquitectura de la red, el número de épocas y el tamaño del *batch*. Los valores clave registrados han sido la pérdida del discriminador, su precisión y la pérdida del generador.

5.1.1.1 Evaluación cuantitativa: precisión del discriminador.

La Figura 5.1 muestra la evolución de la precisión del discriminador para la resolución de 28×28 . Aunque el entrenamiento parte de una precisión intermedia (53.12%), se alcanza rápidamente un rendimiento por encima del 90% en las primeras 200 épocas. Sin embargo, a partir de la época 600, se evidencia una caída progresiva que sugiere síntomas de sobreajuste o pérdida de estabilidad del modelo.

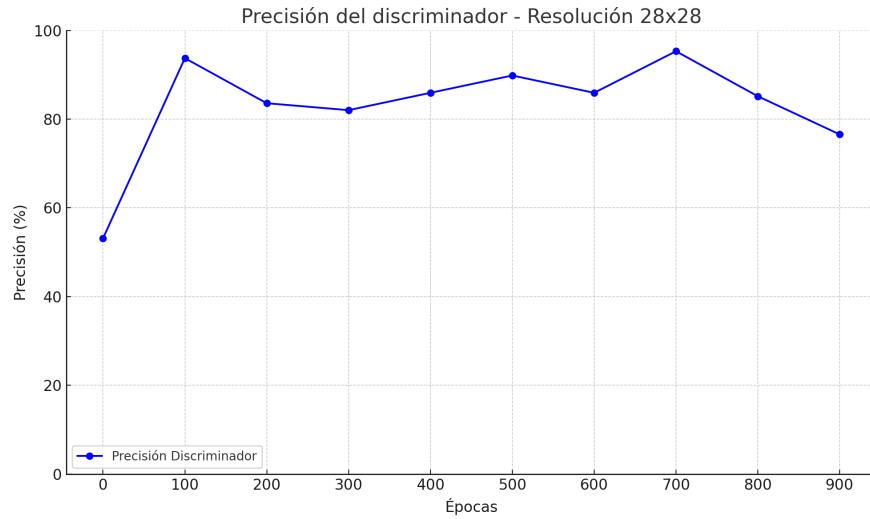


Figura 5.1: Precisión del discriminador - Resolución 28x28.

En el caso de 52×52 , representado en la Figura 5.2, el rendimiento ha sido más estable a lo largo del tiempo. El discriminador alcanza una precisión del 99.22% en la época 100, y aunque se producen fluctuaciones, la precisión se mantiene por encima del 85% en la mayoría de las iteraciones, alcanzando un máximo del 96.09% hacia el final del entrenamiento.

No obstante, se observa una caída abrupta en la época 500, donde la precisión del discriminador desciende temporalmente a su valor mínimo. Este descenso puede deberse a un repunte en la calidad de las imágenes generadas por el generador en ese momento del entrenamiento, lo que dificulta temporalmente la tarea del discriminador y provoca un descenso en su capacidad de clasificación. Este tipo de oscilaciones es característico de las redes adversariales, en las que el progreso de un modelo afecta directamente al comportamiento del otro. Tras esta caída, el discriminador logra adaptarse de nuevo y recuperar su precisión en las siguientes épocas.

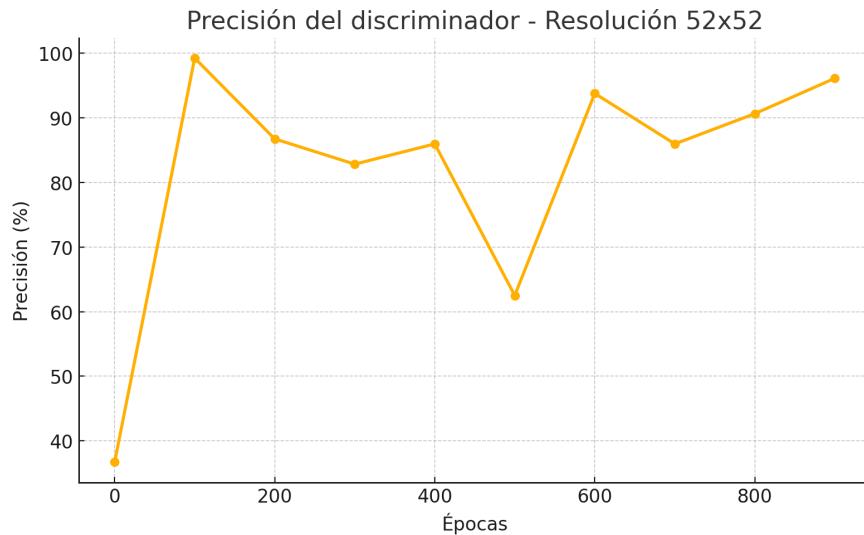


Figura 5.2: Precisión del discriminador - Resolución 52x52.

Para la resolución de 96×96 , se han realizado tres ejecuciones independientes, cuyos resultados agregados se presentan en la Figura 5.3. Las tres ejecuciones muestran patrones similares, con una evolución inicial positiva y un mantenimiento general por encima del 90% durante gran parte del entrenamiento. Las pequeñas diferencias entre ejecuciones sugieren una ligera sensibilidad a la inicialización aleatoria o a los datos seleccionados en cada *batch*.

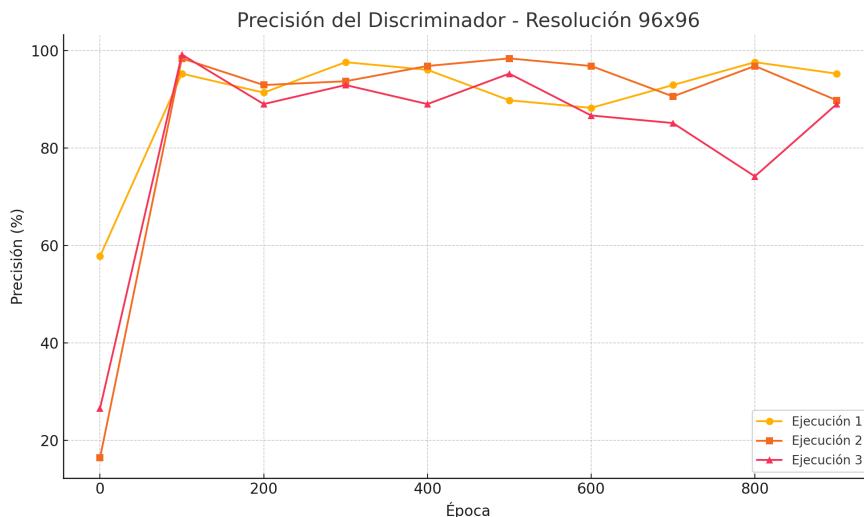


Figura 5.3: Comparativa de precisión entre tres ejecuciones independientes con resolución 96x96.

La comparativa general entre las tres resoluciones puede observarse en la Figura 5.4. Se aprecia que las resoluciones medias y altas tienden a ofrecer un mejor rendimiento global, tanto por precisión como por estabilidad, frente a la resolución más baja, que muestra una caída sostenida en las últimas fases del entrenamiento.

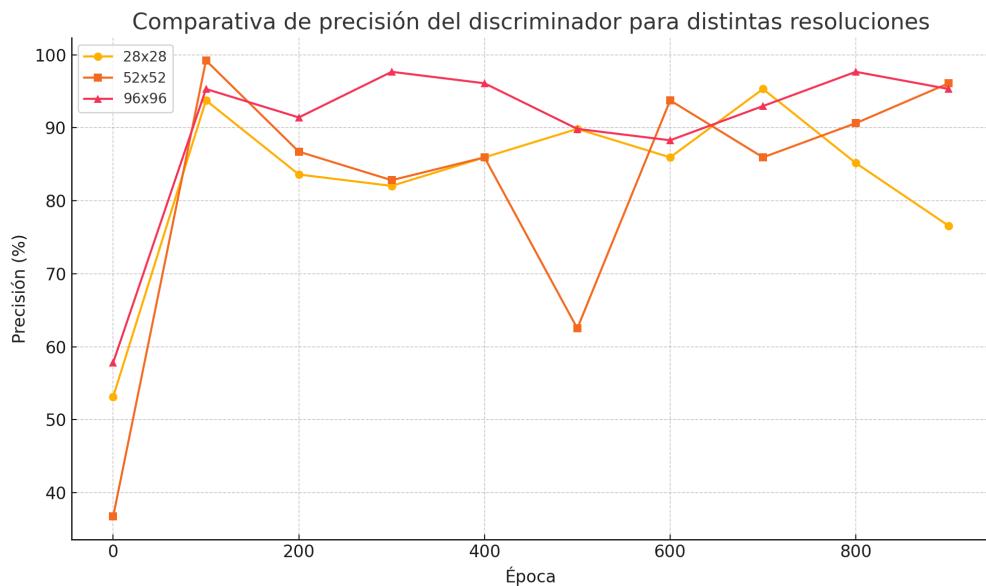


Figura 5.4: Comparativa de precisión para las tres resoluciones evaluadas.

5.1.1.2 Evaluación cualitativa: muestras generadas.

Durante el entrenamiento con imágenes de 96×96 , se han generado muestras visuales cada 100 épocas. La Figura 5.5 recoge ejemplos correspondientes a las épocas 0, 400 y 900. Se observa una mejora significativa en la definición facial, pasando de formas abstractas e indistintas a rostros con una estructura más clara y coherente. Esto sugiere que, si bien el generador no siempre mejora linealmente en pérdida, sí lo hace a nivel perceptual.

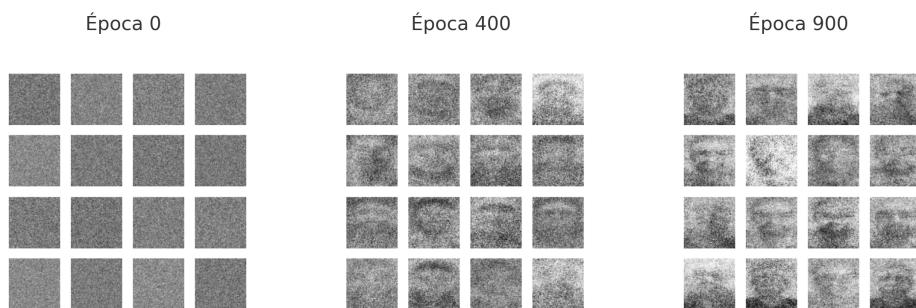


Figura 5.5: Imágenes generadas por la GAN en distintas épocas (resolución 96x96).

En conjunto, los resultados indican que resoluciones más altas, aunque computacionalmente más exigentes, proporcionan una base más rica para que el modelo aprenda patrones visuales complejos. A pesar de una mayor sensibilidad a los hiperparámetros, estas configuraciones permiten obtener mejores resultados tanto en precisión como en calidad de las muestras

generadas. No obstante, se ha identificado que el control del sobreajuste y la estabilidad durante el entrenamiento se vuelve más crítico conforme aumenta la resolución.

5.1.2 Etapa 3: Evaluación con datos de emociones (gan3.py)

En esta tercera etapa, se ha entrenado un modelo GAN utilizando el conjunto de datos FER2013, el cual contiene rostros en escala de grises etiquetados con expresiones emocionales como felicidad, tristeza, enfado o sorpresa. Las imágenes se han redimensionado a 48x48 píxeles, conservando suficiente detalle facial para que el modelo pueda capturar patrones emocionales representativos, a la vez que se ha reducido el coste computacional.

El script `gan3.py` se ha ejecutado tres veces de forma independiente con el objetivo de evaluar la estabilidad del entrenamiento y la calidad de las imágenes generadas. Durante las primeras épocas, se ha observado que el discriminador ha alcanzado niveles elevados de precisión, indicando que ha sido capaz de diferenciar con claridad entre imágenes reales y generadas. Sin embargo, conforme ha avanzado el entrenamiento y el generador ha mejorado su capacidad de producir muestras más realistas, la precisión del discriminador ha disminuido gradualmente. Esta evolución es coherente con el comportamiento esperado en redes generativas adversarias.

En la Figura 5.6 se representa la evolución de la precisión del discriminador para las tres ejecuciones. Se puede observar que, aunque los valores presentan oscilaciones a lo largo del entrenamiento, todas han seguido una tendencia similar, con un pico de precisión temprano seguido de una estabilización por debajo del 70%.

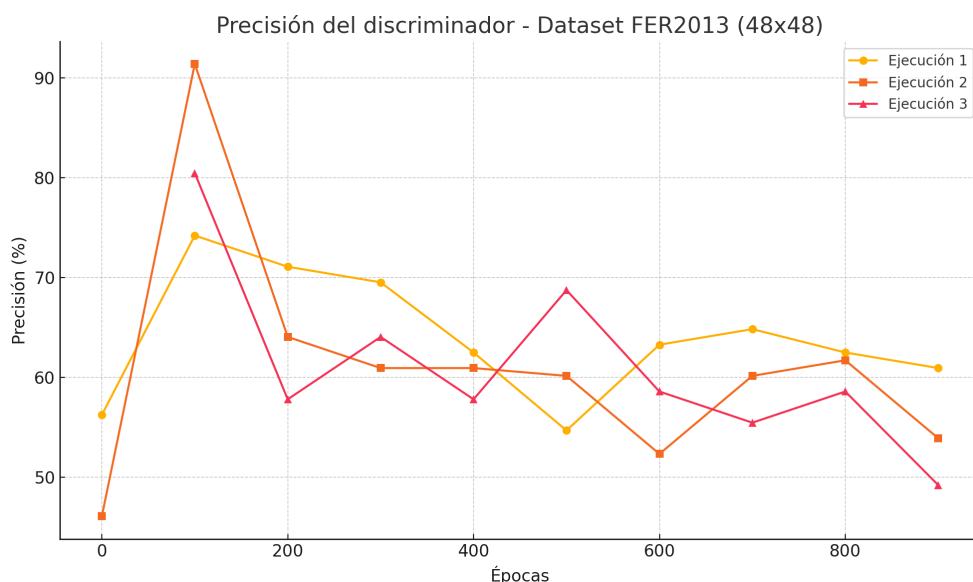


Figura 5.6: Evolución de la precisión del discriminador para las tres ejecuciones del script `gan3.py` utilizando el dataset FER2013 (48x48).

En lo que respecta al generador, se ha percibido una mejora progresiva en la calidad y definición de las imágenes a medida que han transcurrido las épocas. Las muestras generadas

en las primeras fases han presentado ruido aleatorio sin estructura facial discernible, mientras que hacia la mitad del entrenamiento han comenzado a aparecer contornos y siluetas de rostros. En las últimas épocas, los rostros generados han mostrado rasgos cada vez más definidos, incluyendo ojos, nariz y boca, aunque con ciertas imperfecciones y artefactos.

En la Figura 5.7 se muestra una recopilación visual del estado de las imágenes generadas desde la época 0 hasta la 900, en intervalos de 100 épocas. Esta progresión ilustra de manera clara la transición desde patrones aleatorios hasta composiciones faciales reconocibles.

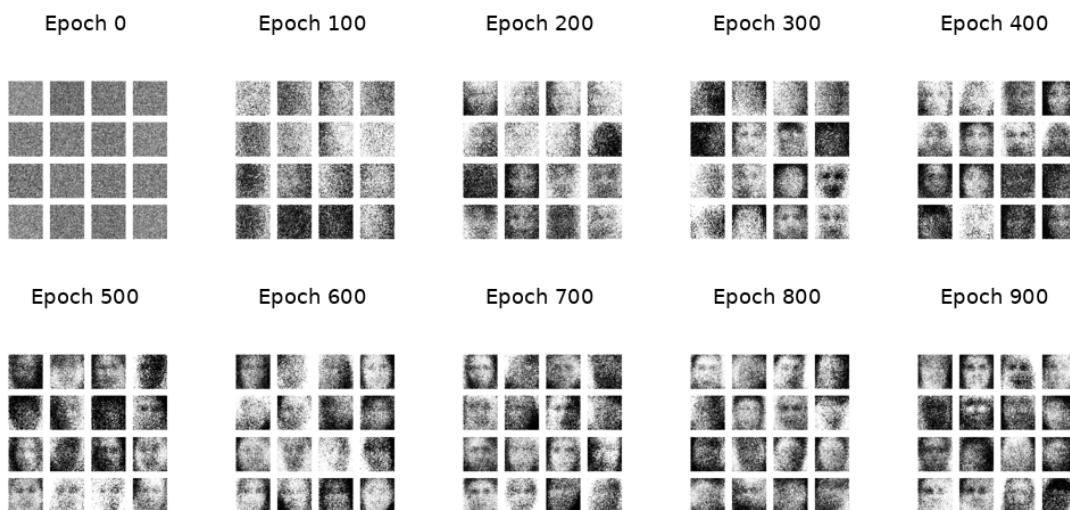


Figura 5.7: Evolución de las imágenes generadas por `gan3.py` desde la época 0 hasta la 900. Puede apreciarse una mejora notable en la estructura facial y la coherencia visual.

Con el objetivo de resaltar dicha evolución, se ha realizado una selección de tres momentos clave del entrenamiento: las épocas 0, 400 y 900. Esta selección se presenta en la Figura 5.8, donde se aprecia claramente el progreso del generador desde una distribución aleatoria de ruido hasta la generación de rostros reconocibles con expresiones faciales diferenciadas.

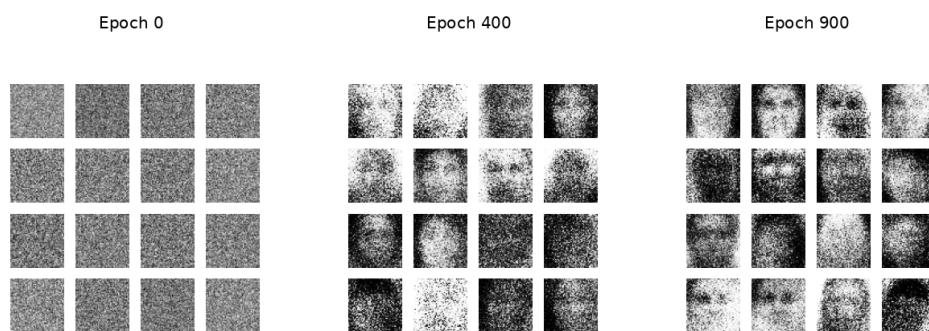


Figura 5.8: Comparativa visual de las imágenes generadas por el modelo entrenado con `gan3.py` en las épocas 0, 400 y 900. Se observa un incremento progresivo en la calidad de los rostros sintetizados.

Aunque la Figura 5.8 proporciona una comparativa general entre las épocas 0, 400 y 900, la escala de la imagen limita la apreciación de ciertos detalles faciales relevantes. Por este motivo, a continuación se presentan las tres muestras de forma individual y ampliada para facilitar una observación más precisa del progreso visual alcanzado por el generador.

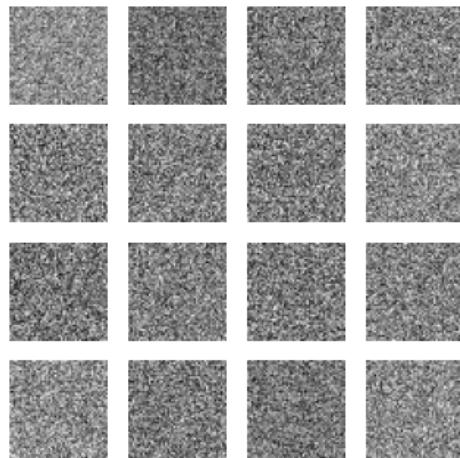


Figura 5.9: Imágenes generadas en la época 0 con `gan3.py`. Se observa ruido aleatorio sin rasgos definidos.

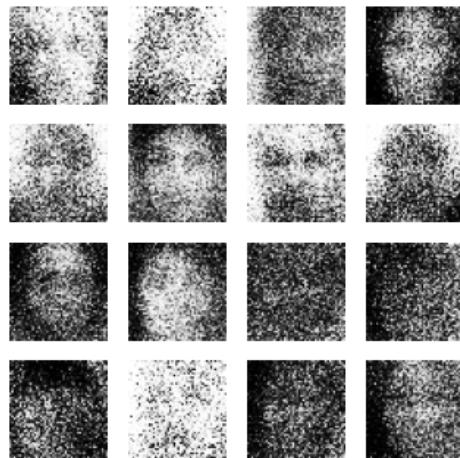


Figura 5.10: Imágenes generadas en la época 400 con `gan3.py`. Comienzan aemerger estructuras faciales básicas.

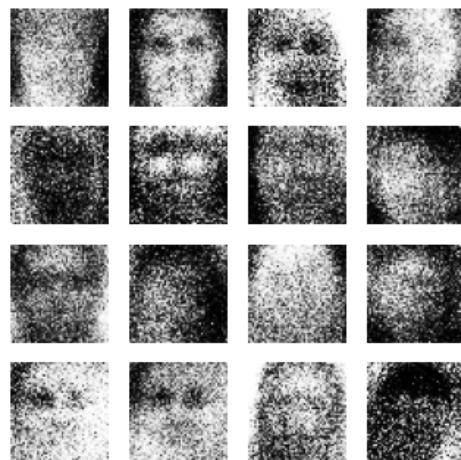


Figura 5.11: Imágenes generadas en la época 900 con `gan3.py`. Se distinguen claramente rostros con expresiones faciales.

En conjunto, los resultados han demostrado que la arquitectura empleada ha sido capaz de aprender representaciones visuales coherentes a partir de un conjunto de datos emocionalmente etiquetado. No se ha detectado colapso de modo, ya que el modelo ha generado imágenes variadas entre sí, y la evolución de la calidad ha sido consistente entre las distintas ejecuciones. Las diferencias observadas entre entrenamientos pueden atribuirse a la aleatoriedad inherente al proceso de inicialización y optimización.

5.1.3 Etapa 4: Evaluación cruzada con conjunto de test (`gan4.py`)

En esta etapa, se ha evaluado la capacidad de generalización del modelo GAN utilizando un conjunto de test independiente, manteniéndose así una separación estricta respecto a los datos de entrenamiento. Esta evaluación se ha realizado mediante el script `gan4.py`, aplicando el modelo entrenado con FER2013 (48x48) sobre muestras no vistas durante el proceso de entrenamiento.

Se han llevado a cabo tres ejecuciones distintas para observar el comportamiento del discriminador frente a datos externos. En todas ellas se ha registrado una tendencia similar: tras una primera fase con alta precisión en las primeras épocas, se ha evidenciado una oscilación progresiva y una ligera pérdida de rendimiento, especialmente a partir de la época 500. Este fenómeno puede estar relacionado con un sobreajuste progresivo del discriminador a las características del conjunto de entrenamiento, lo que le dificulta identificar imágenes reales en el conjunto de test.

En la Figura 5.12 se muestra la evolución de la precisión del discriminador sobre el conjunto de test para las tres ejecuciones realizadas. Se aprecia que, aunque el comportamiento inicial ha sido robusto (superando en algunos casos el 80% de precisión), con el tiempo se ha observado una reducción sostenida del rendimiento, con caídas hasta valores entre el 50% y 60%.

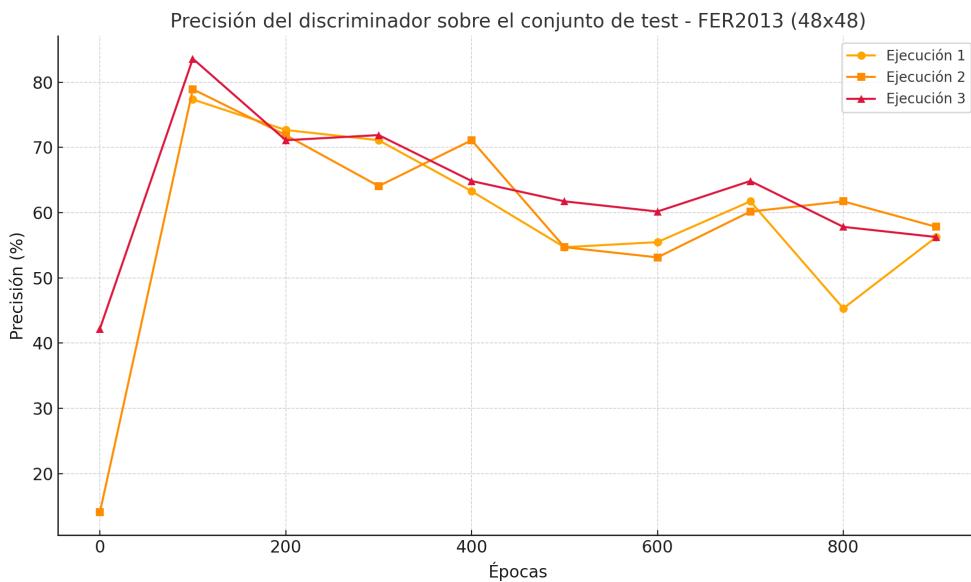


Figura 5.12: Precisión del discriminador sobre el conjunto de test en las tres ejecuciones de `gan4.py`.

Este comportamiento se puede interpretar como una señal de sobreajuste del modelo, cuya capacidad de discriminación mejora durante las primeras etapas del entrenamiento pero posteriormente tiende a adaptarse demasiado a las características específicas del conjunto de entrenamiento. Como complemento, se ha analizado la distribución de imágenes en el conjunto de entrenamiento (Figura 5.13), así como su comparación directa con el conjunto de test (Figura 5.14). Estas visualizaciones permiten contextualizar las diferencias y similitudes entre ambos subconjuntos y ofrecen una explicación visual sobre la posible pérdida de generalización detectada.

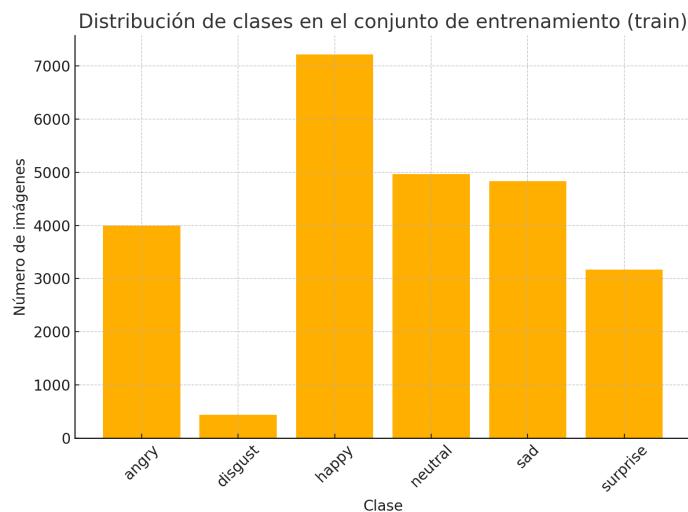


Figura 5.13: Distribución de las clases emocionales en el conjunto de entrenamiento (FER2013).

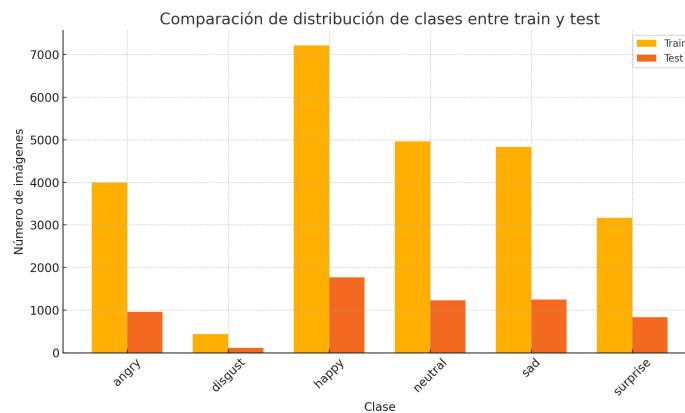


Figura 5.14: Comparativa visual entre muestras del conjunto de entrenamiento y del conjunto de test.

Esta caída del rendimiento, acompañada de una estabilización en niveles bajos de precisión, sugiere que el discriminador ha terminado por memorizar patrones específicos del conjunto de entrenamiento en lugar de aprender características generalizables. Esta interpretación se ve reforzada al observar que, pese a la distribución equilibrada entre clases, pueden existir diferencias sutiles (como expresividad, resolución o calidad visual) entre las imágenes de *train* y *test* que el modelo no ha sabido abstraer.

Por tanto, esta fase ha servido para evidenciar que, a pesar de una buena capacidad inicial de generalización, el modelo tiende a sobreajustarse conforme avanza el entrenamiento. La evaluación cruzada ha permitido identificar este fenómeno y abre la posibilidad de explorar mecanismos de regularización o técnicas de parada temprana para mitigar sus efectos en futuras iteraciones.

5.1.4 Consideraciones globales sobre el aprendizaje adversarial

A lo largo de las distintas etapas experimentales se han identificado una serie de comportamientos comunes que permiten establecer conclusiones generales sobre el funcionamiento del sistema GAN. En primer lugar, se ha confirmado que el proceso de entrenamiento está marcado por una dinámica adversarial compleja, donde el equilibrio entre generador y discriminador resulta esencial para evitar problemas como el colapso de modo o la pérdida inestable.

Uno de los riesgos más relevantes observados ha sido el *colapso de modo*, fenómeno que se manifiesta cuando el generador deja de producir muestras diversas y comienza a generar únicamente un conjunto muy reducido de imágenes, generalmente similares entre sí. Esto suele ocurrir cuando el generador descubre un patrón que logra engañar al discriminador de forma consistente, y lo explota de forma reiterada, en lugar de aprender la distribución completa del conjunto de datos reales. En consecuencia, aunque las imágenes generadas puedan parecer visualmente correctas, el modelo pierde su capacidad de representar la variedad original del conjunto, lo que limita su utilidad y su capacidad de generalización.

Por otro lado, se ha evidenciado que el sobreajuste del discriminador es una amenaza real, especialmente en fases avanzadas del entrenamiento. Este fenómeno se ha observado en

varias ejecuciones, donde la precisión del discriminador sobre el conjunto de entrenamiento se ha mantenido alta, mientras que su rendimiento sobre datos no vistos ha disminuido significativamente. Esta desconexión entre ambos comportamientos indica una especialización excesiva en las muestras conocidas, dificultando la detección efectiva de imágenes generadas en contextos diferentes.

La pérdida adversarial también ha mostrado una evolución oscilante en numerosos casos, reflejando la sensibilidad del sistema a factores como la inicialización aleatoria, el tamaño del *batch* o la tasa de aprendizaje. Esta inestabilidad hace necesaria la monitorización continua del proceso de entrenamiento mediante métricas cuantitativas y cualitativas.

Para mitigar estos problemas, se han implementado diversas estrategias de control y validación. Por un lado, se han generado muestras visuales periódicas (cada 100 épocas) que permiten una inspección manual del progreso del generador. Esta validación visual ha resultado fundamental para detectar desviaciones o anomalías que no siempre se reflejan en las métricas numéricas. Por otro lado, se ha considerado la posibilidad de aplicar técnicas como la parada temprana (*early stopping*) en futuras versiones del sistema, con el fin de evitar que el modelo continúe entrenando más allá del punto óptimo y empiece a sobreajustarse o degradar su rendimiento.

En conjunto, estos hallazgos refuerzan la idea de que el aprendizaje adversarial, si bien ofrece un marco potente para la generación de datos sintéticos, implica también una elevada complejidad en su gestión y evaluación. La identificación de patrones comunes, como el colapso de modo o el sobreajuste, junto con la utilidad de mecanismos de validación intermedia, constituye una base sólida sobre la que seguir construyendo. Este análisis transversal permite comprender mejor las limitaciones prácticas del sistema actual y proporciona un contexto fundamentado para interpretar de forma crítica los resultados obtenidos en cada una de las etapas previas.

5.2 Evaluación del Análisis de Profundidad con MiDaS

Con el objetivo de analizar diferencias estructurales entre imágenes reales y copias impresas, se ha empleado el modelo MiDaS para la generación de mapas de profundidad. MiDaS, basado en redes neuronales profundas, permite estimar la distancia relativa de cada píxel respecto a la cámara a partir de una única imagen. Esta propiedad resulta especialmente útil en escenarios de autenticación biométrica, donde los ataques por presentación pueden introducir variaciones sutiles en la geometría de la escena.

El análisis se ha llevado a cabo sobre conjuntos de imágenes reales y sus correspondientes impresiones fotocopiadas, procesadas en parejas para facilitar la comparación directa de sus características de profundidad. Para automatizar el flujo de trabajo, se han desarrollado y utilizado scripts específicos como `run_midas.py`, encargado de generar los mapas de profundidad, y `analyze.py`, orientado al análisis y extracción de características de las regiones de interés (ROI). Estos scripts permiten el procesamiento en lote de múltiples imágenes, garantizando la reproducibilidad de los experimentos y reduciendo posibles sesgos manuales.

Dado el carácter preliminar y exploratorio del estudio, se ha optado por un enfoque cualitativo. No se han definido umbrales de clasificación fija ni métricas cerradas; en su lugar, se ha priorizado la observación sistemática de patrones visuales y la evaluación de tendencias generales en las texturas y distribuciones de profundidad. Este planteamiento flexible ha per-

mitido identificar diferencias relevantes que podrán ser formalizadas y explotadas en trabajos posteriores orientados al desarrollo de sistemas automáticos de detección de suplantación.

5.2.1 Mapas de Profundidad y Visualización Directa

Con el fin de evaluar visualmente las diferencias de profundidad entre imágenes reales y suplantadas mediante fotocopia, se ha realizado un análisis cualitativo utilizando los mapas generados con MiDaS.

Durante el proceso experimental, se han recopilado numerosas imágenes de diferentes individuos, capturadas en diversas condiciones de iluminación y entorno. Cada persona ha sido registrada en varias tomas distintas, tanto en imagen original como en su respectiva versión impresa y posteriormente fotografiada de nuevo.

Para esta sección, se han seleccionado cuatro ejemplos representativos del conjunto completo, correspondientes a las muestras denominadas **Ester5**, **Pepe3**, **Ester2** (en exteriores) y **Pepe1**. La numeración asociada a cada nombre refleja la posición de la captura dentro de la serie general obtenida para cada sujeto. La selección busca ilustrar las principales tendencias observadas en el análisis de profundidad, permitiendo extraer conclusiones generales sobre el comportamiento de las imágenes reales frente a las suplantadas.

Caso 1: Ester5 (interior)

En la imagen real de Ester5, el mapa de profundidad presenta una correcta diferenciación entre el rostro y el fondo, con transiciones graduales en las zonas de la cabeza, el cuello y la pizarra situada detrás. En contraste, la imagen de fotocopia muestra un mapa considerablemente más homogéneo: el rostro pierde variaciones de profundidad y las zonas que deberían mostrar relieve aparecen aplanadas. La pérdida de texturas finas es clara en la parte de la ropa y los contornos de los hombros.

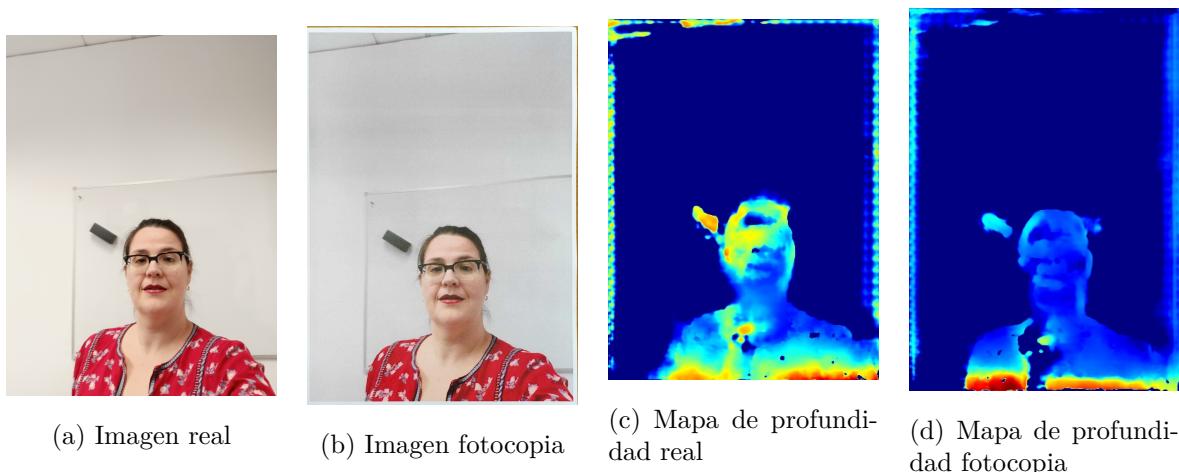


Figura 5.15: Comparativa entre imagen real y fotocopia en el caso **ester5**.

Caso 2: Pepe3 (interior)

El mapa de profundidad real de Pepe3 muestra variaciones suaves en el rostro, el fondo y el brazo, reflejando correctamente las diferencias de distancia al sensor. En cambio, en la versión fotocopiada, la profundidad se concentra en bloques más uniformes, perdiendo la riqueza de detalles locales. Las zonas de sombra y la textura de la camiseta, visibles en la imagen real, desaparecen casi por completo en la fotocopia.

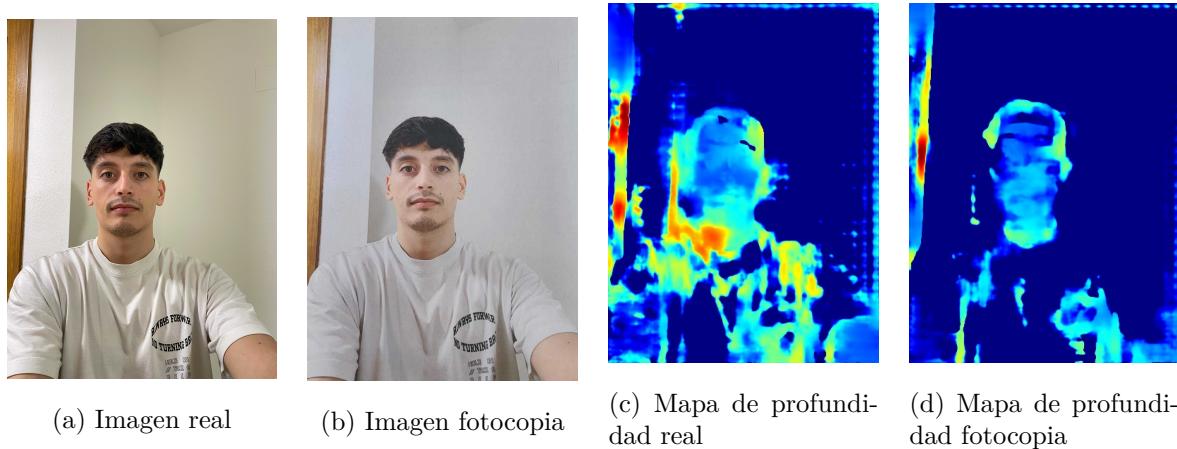


Figura 5.16: Comparativa entre imagen real y fotocopia en el caso **pepe3**.

Caso 3: Ester2 (exterior)

En condiciones exteriores, como en el caso de Ester2, el mapa de profundidad real refleja la riqueza de detalles del entorno, destacando claramente el fondo (árboles, edificio) respecto al sujeto. En la imagen de la fotocopia, aunque el fondo conserva cierto nivel de variación —heredado de la riqueza visual de la escena original—, el rostro aparece significativamente aplastado, con una pérdida evidente de estructura tridimensional. Esta reducción de relieve es especialmente visible en las zonas de la nariz y los pómulos, lo que sugiere una representación superficial propia de una copia impresa. Este caso evidencia que en escenarios exteriores con alta complejidad visual, la suplantación mediante fotocopia puede ser más difícil de detectar únicamente por análisis del fondo.

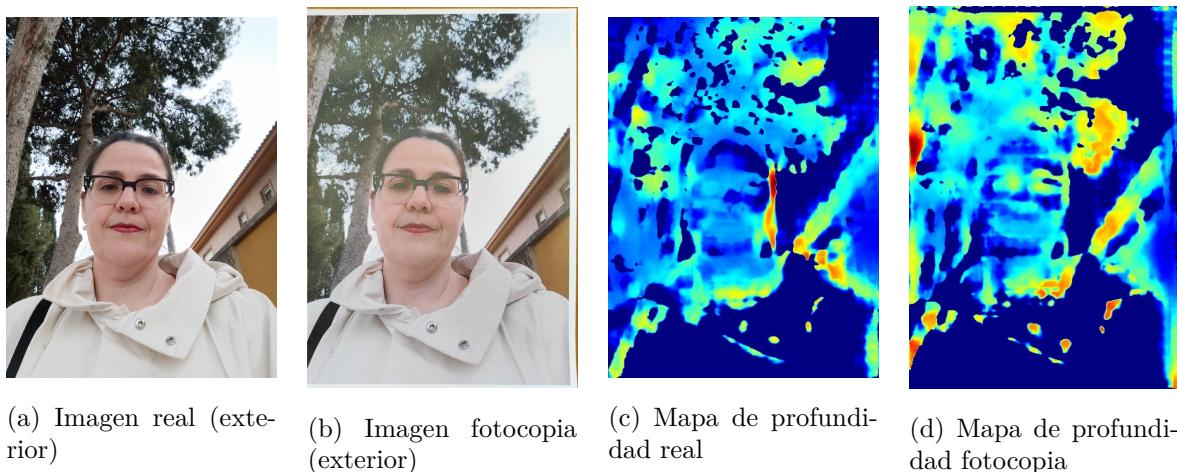


Figura 5.17: Comparativa entre imagen real y fotocopia en el caso **ester2** (exterior).

Caso 4: Pepe1 (interior)

Finalmente, el caso de Pepe1 muestra un comportamiento consistente con los anteriores ejemplos interiores. En la imagen real, los contornos del rostro, la pared y la puerta son bien diferenciados en términos de profundidad. En la fotocopia, en cambio, las transiciones desaparecen, el rostro se aplana considerablemente y el fondo se vuelve más homogéneo.

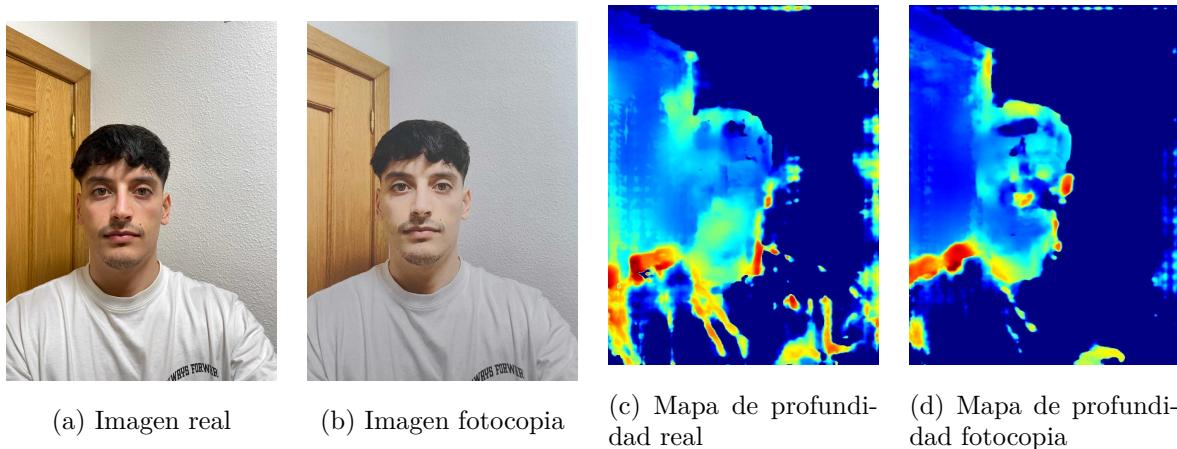


Figura 5.18: Comparativa entre imagen real y fotocopia en el caso **pepe1**.

De los casos analizados se desprenden varias observaciones clave:

- La fotocopia reduce de forma consistente la riqueza de detalles de profundidad, aplanando superficies que en imágenes reales presentan variaciones tridimensionales sutiles.
- La pérdida de estructura es especialmente visible en zonas faciales y en el contacto sujeto-fondo.

- En entornos exteriores, los detalles del fondo pueden dificultar la detección de suplantaciones si no se focaliza el análisis en el rostro.
- Factores como iluminación, resolución de impresión y calidad de captura tienen un impacto directo en la calidad del mapa de profundidad obtenido.

Estos resultados refuerzan la utilidad del análisis de mapas de profundidad como método preliminar para detectar intentos de suplantación en sistemas biométricos, aunque su integración debería considerar una combinación de múltiples estrategias para una robustez completa.

5.2.2 Histogramas de Profundidad y Análisis de Textura

Además del análisis visual de los mapas de profundidad, se ha realizado un estudio estadístico de la distribución de los valores de profundidad presentes en las imágenes. Para ello, se han generado histogramas, que permiten observar gráficamente cómo se reparten los valores de profundidad estimados por el modelo MiDaS en cada imagen.

Un histograma de profundidad representa en el eje horizontal los valores de distancia relativa estimados, y en el eje vertical la frecuencia de aparición de esos valores en la imagen. De esta manera, es posible analizar si los valores de profundidad están dispersos (lo que refleja una estructura tridimensional rica) o si están concentrados (lo que indica superficies más planas).

En este estudio se han considerado dos enfoques:

- Generación de histogramas sobre la imagen completa de profundidad.
- Generación de histogramas centrados únicamente en la región facial (ROI).

Ambas estrategias permiten evaluar diferencias en la dispersión de los valores de profundidad entre imágenes reales y suplantadas mediante fotocopia.

5.2.3 Histogramas de la Imagen Completa

En una primera fase, se han generado histogramas considerando la imagen completa, es decir, incluyendo tanto el rostro como el fondo de la escena.

Con el objetivo de complementar el análisis visual presentado en las Figuras 5.18 y 5.15, se han generado histogramas de profundidad correspondientes a dos casos representativos: **Pepe1** y **Ester5**. En cada caso, se comparan los valores de profundidad de las imágenes reales y sus respectivas versiones fotocopiadas, permitiendo observar de forma cuantitativa cómo varía la distribución estructural.

Caso 1: Pepe1 (interior)

En la imagen real del caso Pepe1 (véase Figura 5.18), el mapa de profundidad presentaba un amplio rango de valores y transiciones suaves entre diferentes planos. El histograma mostrado en la Figura 5.19 confirma esta observación: en la imagen original se observa una alta concentración inicial (valores bajos), seguida de una distribución progresiva que se extiende hasta niveles altos de profundidad.

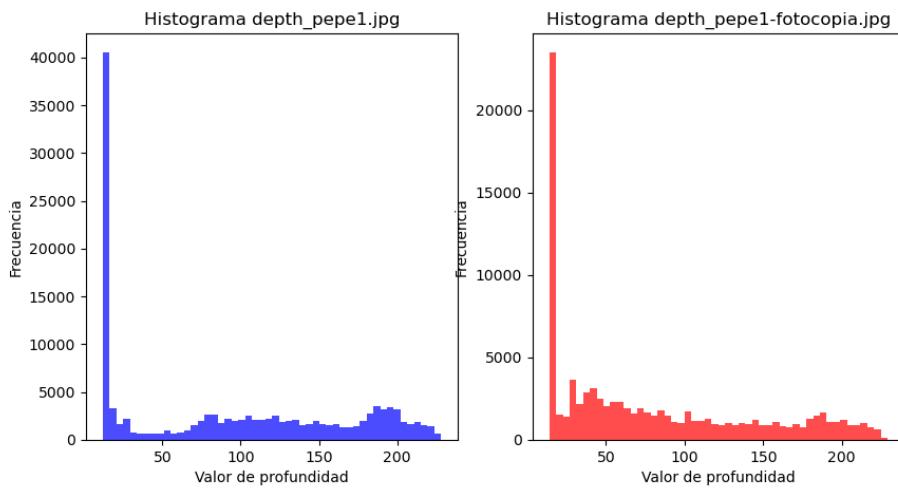


Figura 5.19: Histogramas de profundidad para la imagen real y la fotocopia en el caso **pepe1**.

En cambio, la imagen fotocopiada presenta una clara concentración de valores en el primer rango del histograma, con una caída brusca de frecuencia a partir de los 50 niveles de profundidad. Esta distribución más estrecha indica una pérdida de variabilidad espacial, coherente con la naturaleza bidimensional de una imagen impresa. La diferencia visual entre ambos histogramas pone de manifiesto una menor riqueza tridimensional en la fotocopia.

Caso 2: Ester5 (interior)

De forma similar, en el caso **Ester5** —ver Figura 5.15— se observa una clara distinción entre imagen real y fotocopia, reflejada también en los histogramas de la Figura 5.20.

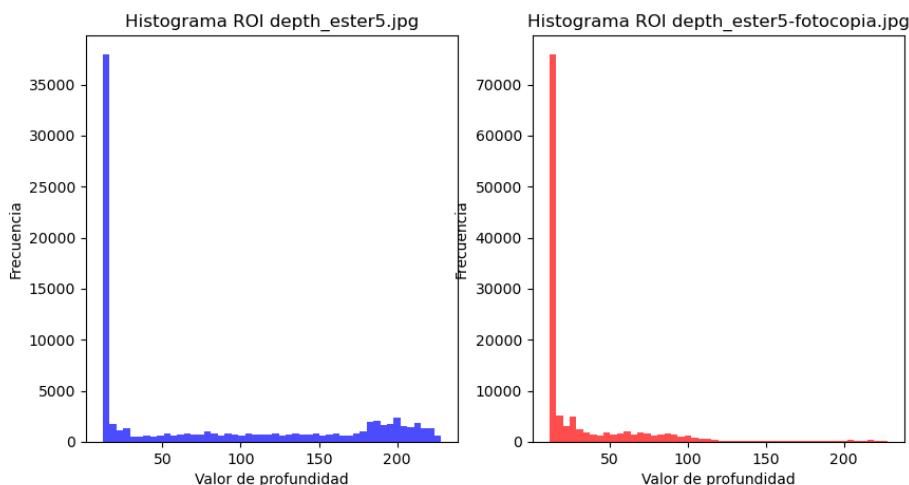


Figura 5.20: Histogramas de profundidad para la imagen real y la fotocopia en el caso **ester5**.

En la imagen real, los valores de profundidad están más repartidos, incluyendo una mayor frecuencia en el rango alto (entre 150 y 300). Este patrón sugiere la existencia de variaciones estructurales tanto en el rostro como en el entorno. En cambio, la imagen fotocopiada muestra una enorme acumulación de valores entre 0 y 100, con apenas contribución más allá de ese umbral. Esto refleja una notable planitud, en la que la mayoría de los píxeles se encuentran en una profundidad casi constante.

Por tanto, ambos histogramas respaldan cuantitativamente lo observado en los mapas de profundidad. Las imágenes reales presentan una mayor dispersión de niveles, mientras que las fotocopias tienden a una distribución abruptamente decreciente, con la mayor parte de los valores concentrados en los primeros rangos. Estos patrones son indicativos de una estructura más rica en las imágenes auténticas y refuerzan la viabilidad del uso de histogramas como herramienta discriminativa en sistemas automáticos de detección de suplantaciones faciales.

5.2.3.1 Histogramas de la Región de Interés (ROI)

Para afinar el análisis y reducir el impacto del fondo, se han generado histogramas de profundidad exclusivamente sobre la región de interés facial (ROI).

Caso 1: Ester2 (exterior)

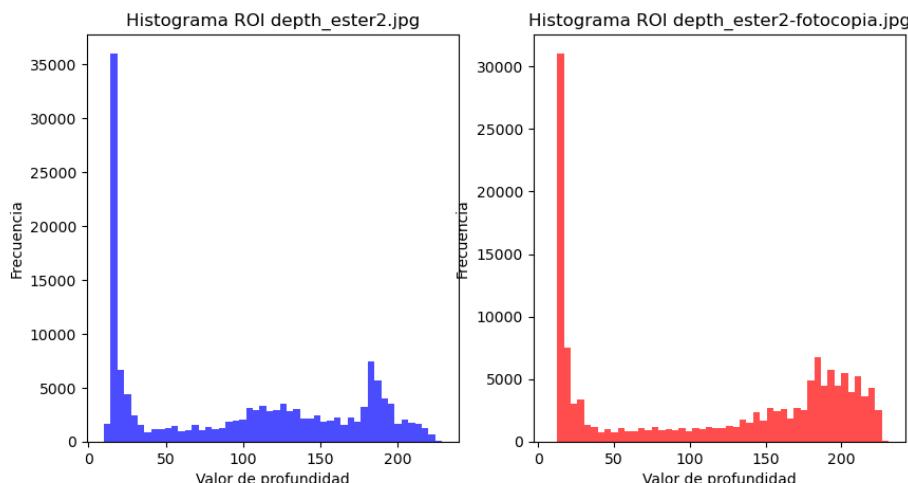


Figura 5.21: Comparativa de histogramas de profundidad en la región facial en el caso **ester2** (exterior).

En este caso capturado en exteriores, la imagen real presenta una mayor dispersión de valores de profundidad en la región facial. Aunque en la fotocopia se mantiene cierta variabilidad debido a la riqueza visual del entorno natural, se observa una reducción clara en la diversidad de profundidades.

Caso 2: Ester4 (interior)

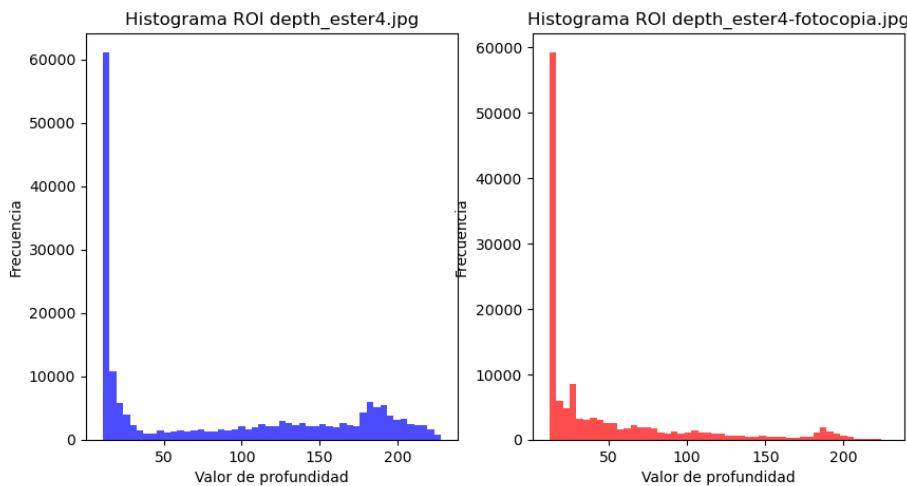


Figura 5.22: Comparativa de histogramas de profundidad en la región facial en el caso **ester4** (inferior).

En interiores, donde el fondo presenta menor variabilidad, la diferencia entre la imagen real y la fotocopia es aún más evidente: la fotocopia muestra un fuerte agrupamiento de valores de profundidad, indicando una mayor pérdida de estructura tridimensional en el rostro respecto a la imagen real.

5.2.3.2 Conclusiones del análisis de histogramas

El análisis conjunto de los histogramas evidencia que las imágenes fotocopiadas tienden a perder riqueza tridimensional, concentrando los valores de profundidad en rangos más estrechos. Este fenómeno es observable tanto en el análisis de imagen completa como, de forma más acentuada, al centrarse únicamente en la región facial (ROI).

La segmentación facial permite un análisis más preciso, ya que elimina el ruido introducido por el fondo y resalta las diferencias específicas en la estructura tridimensional del rostro. Estos resultados refuerzan la idea de que el análisis de histogramas de profundidad constituye una herramienta efectiva para detectar suplantaciones faciales basadas en la presentación de imágenes impresas.

Como posible línea futura de investigación, se plantea desarrollar mecanismos automáticos de segmentación facial previa al análisis de profundidad, optimizando así la fiabilidad de los sistemas de detección de suplantaciones.

5.2.4 Comparación entre Métodos de Extracción de ROI

En esta sección se han comparado dos enfoques diferentes para la extracción de la Región de Interés (ROI) aplicada al análisis de mapas de profundidad: la detección por contornos y la detección facial mediante *Haar cascades*. Ambos métodos se han implementado con el objetivo de delimitar la zona del rostro en imágenes reales y suplantadas, permitiendo estudiar con mayor precisión las características de profundidad presentes en esa región.

Para facilitar la comprensión de las muestras utilizadas, en la Figura 5.23 se presentan las imágenes originales reales y sus correspondientes versiones impresas en los casos **pepe1** y **pepe2**, que se han empleado como base en la comparación posterior:



(a) Real - Pepe1 (b) Fotocopia - Pepe1 (c) Real - Pepe2 (d) Fotocopia - Pepe2

Figura 5.23: Imágenes originales y sus correspondientes fotocopias en los casos **pepe1** y **pepe2**.

5.2.4.1 Método 1: Detección por contornos

Este enfoque se ha basado en la detección del contorno más destacado dentro de la imagen de profundidad. Para ello, se ha aplicado una binarización con umbral adaptativo con el fin de resaltar las estructuras prominentes. Posteriormente, se ha seleccionado el contorno de mayor área, asumiendo que correspondía al rostro. A partir de este contorno se ha generado una máscara que ha permitido recortar la región de análisis.

Este método ha tenido como ventaja principal su independencia de la imagen RGB o de características faciales explícitas, lo que le ha conferido cierta flexibilidad. De hecho, ha funcionado incluso en algunas imágenes impresas donde los métodos tradicionales no han detectado rostros. Sin embargo, también se ha mostrado más sensible a elementos no deseados del fondo —como sombras, marcos de puertas o paredes— y ha generado ROIs más irregulares, especialmente en las fotocopias.

5.2.4.2 Método 2: Detección mediante *Haar cascades*

El segundo enfoque ha utilizado un clasificador *Haar cascade* preentrenado sobre la imagen RGB original para localizar la cara. Una vez detectada, se ha definido un rectángulo que se ha proyectado sobre la imagen de profundidad, extrayendo así la ROI correspondiente.

Este método ha sido más directo, regular y fácilmente replicable. En las imágenes reales ha funcionado con gran fiabilidad. Sin embargo, su precisión ha dependido directamente de la calidad de la imagen RGB. En fotocopias donde la cara ha aparecido difusa, iluminada de forma deficiente o con baja resolución, el detector ha tenido problemas para identificar correctamente la región facial.

5.2.4.3 Comparación Visual: Pepe1

A continuación se presentan las ROIs generadas mediante ambos métodos para el caso **pepe1**, tanto para la imagen original como para su fotocopia:

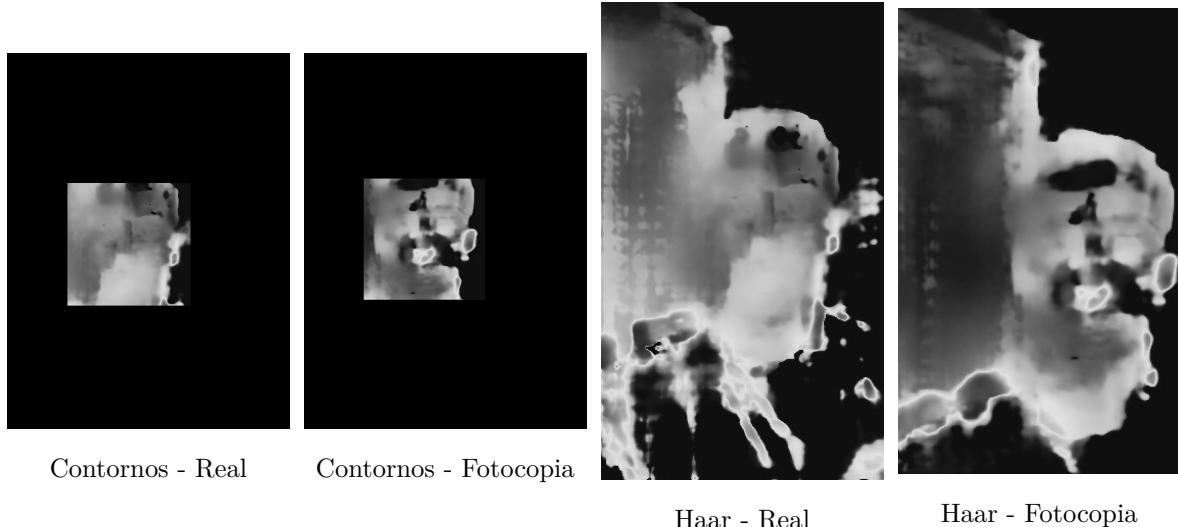


Figura 5.24: Comparativa de ROIs extraídas en el caso **pepe1**.

En este caso, se ha observado que el método por contornos ha generado una ROI deformada en la fotocopia, aunque razonablemente centrada. Por su parte, la detección *Haar* ha ofrecido un recorte más regular y rectangular, aunque ha incluido algo más de fondo, lo cual podría introducir ruido.

5.2.4.4 Comparación Visual: Pepe2

En el caso **pepe2**, los resultados han sido similares, pero con matices relevantes:

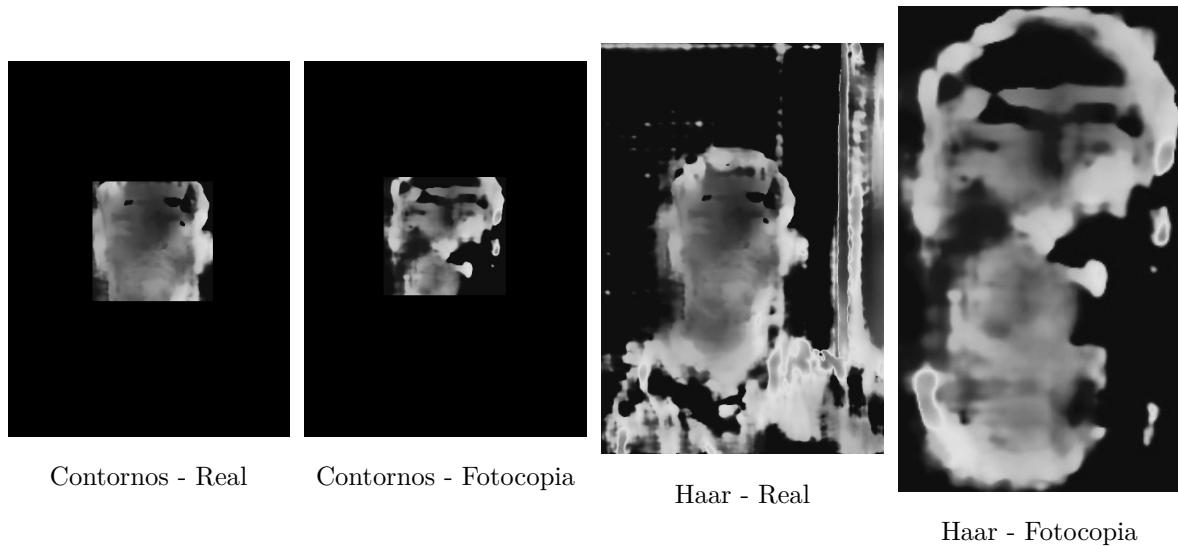


Figura 5.25: Comparativa de ROIs extraídas en el caso **pepe2**.

Se ha comprobado que el método por contornos ha funcionado de forma algo más estable en la imagen real, pero en la fotocopia ha capturado bordes distorsionados que no han pertenecido al rostro. En cambio, la detección facial ha sido más conservadora siendo menos dependiente de las condiciones externas.

Ambos métodos han mostrado ventajas y limitaciones distintas según el tipo de imagen y el estado del rostro. El método basado en *Haar cascades* ha ofrecido resultados más consistentes cuando ha sido posible detectar el rostro correctamente, generando ROIs con forma más regular y centrada. El método por contornos ha sido útil en escenarios donde el rostro no ha sido reconocible con precisión en la imagen RGB, pero ha demostrado una mayor sensibilidad al fondo y a los artefactos de la fotocopia.

Por tanto, no se ha podido afirmar que un método sea universalmente mejor, aunque para este trabajo —donde la regularidad y centrado de la ROI son prioritarios— la técnica de detección facial ha resultado más adecuada en la mayoría de los casos.

5.2.5 Indicadores Cuantitativos y Observaciones Generales

Esta sección presenta un análisis cuantitativo detallado de las imágenes reales y sus correspondientes versiones impresas, a partir de las regiones de interés (ROI) extraídas de los mapas de profundidad generados con el modelo *MiDaS*. El objetivo ha sido caracterizar numéricamente las diferencias estructurales y texturales entre ambas clases de imágenes, con el fin de establecer patrones consistentes que puedan ser útiles en futuros sistemas automáticos de detección de suplantaciones.

Nota sobre el conjunto de datos: Todas las imágenes analizadas en este apartado han sido capturadas en condiciones de interior. Se han excluido explícitamente las imágenes tomadas en exteriores, ya que estas presentaban mayores variaciones de iluminación ambiental, reflejos y fondos no controlados, lo que afectaba negativamente a la estabilidad de los mapas de

profundidad generados. El análisis se ha limitado, por tanto, a un entorno controlado que permite una comparación más fiable entre imágenes reales y sus respectivas fotocopias.

5.2.5.1 Métricas empleadas

Sobre cada ROI facial se han extraído las siguientes métricas:

- **Variación de profundidad:** definida como la diferencia entre el valor máximo y el mínimo de la profundidad dentro de la ROI. Mide el rango estructural tridimensional del rostro.
- **Textura (varianza del Laplaciano):** representa la riqueza de detalles estructurales, útil para diferenciar superficies planas (fotocopias) de rostros reales.
- **Desviación típica de profundidad:** refleja la dispersión de los valores de profundidad y sirve como complemento al rango, aportando robustez frente a valores atípicos.

5.2.5.2 Visualización comparativa

Las siguientes figuras muestran la comparación directa entre imágenes reales y fotocopias para cada una de las métricas mencionadas. Los valores han sido representados para todos los casos analizados:

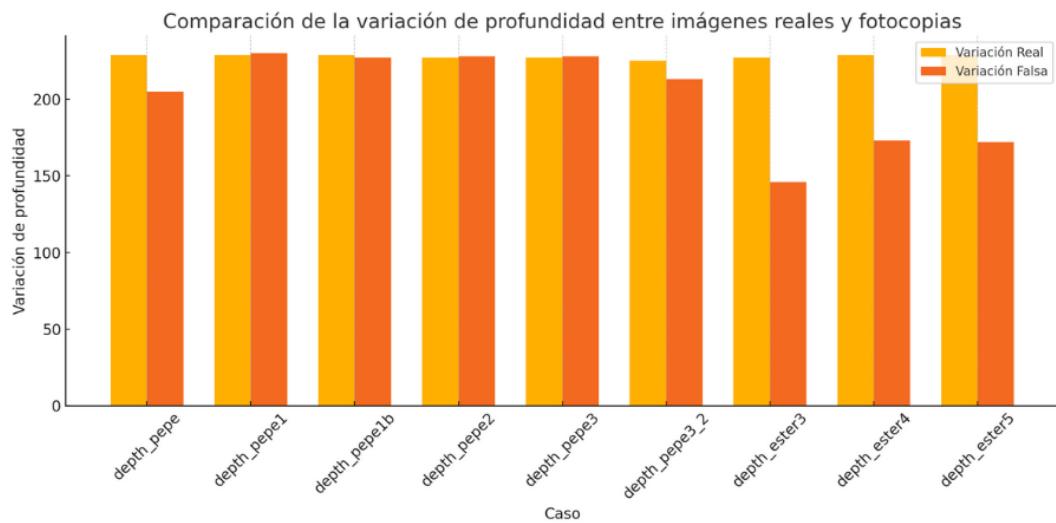


Figura 5.26: Comparación de la variación de profundidad entre imágenes reales y fotocopias.

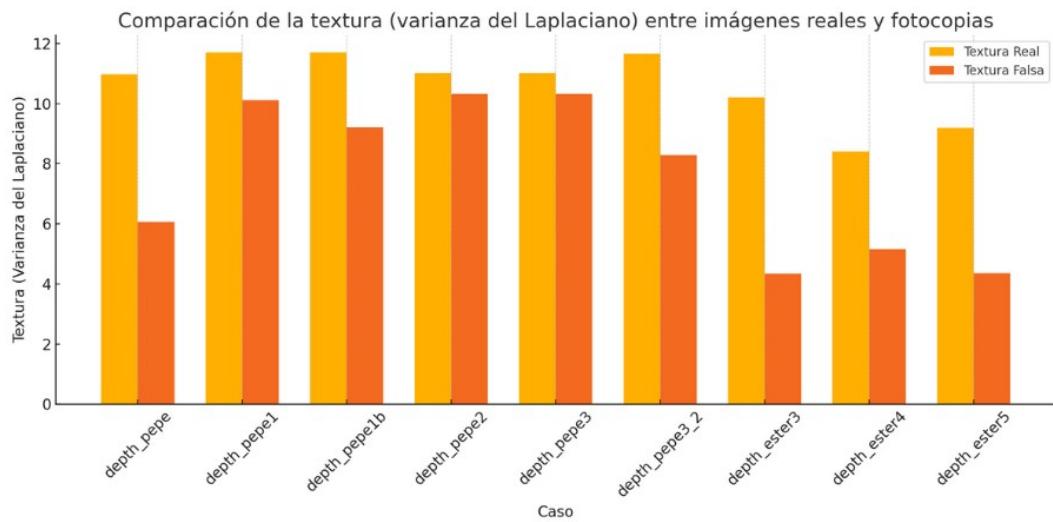


Figura 5.27: Comparación de la textura (varianza del Laplaciano) entre imágenes reales y photocopies.

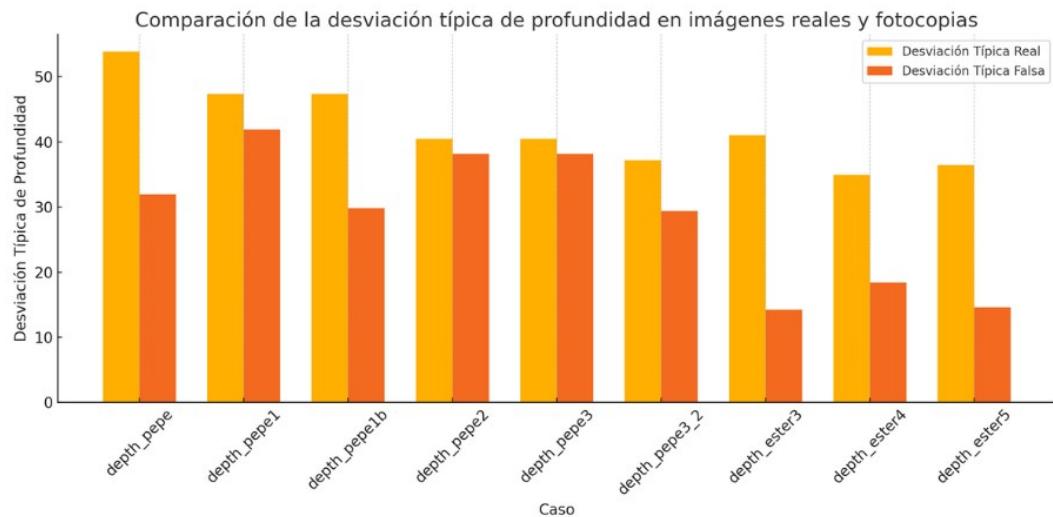


Figura 5.28: Comparación de la desviación típica de profundidad entre imágenes reales y photocopies.

5.2.5.3 Resumen numérico de resultados

Caso	Variación de Profundidad	Textura (Laplaciano)	Desviación Típica
Pepe1	229	11.70	47.37
Pepe1 Fotocopia	230	10.11	41.87
Pepe2	227	11.02	40.44
Pepe2 Fotocopia	228	10.32	38.15
Pepe3	225	11.66	37.21
Pepe3 Fotocopia	213	8.28	29.39
Ester3	227	10.20	40.99
Ester3 Fotocopia	146	4.34	14.23
Ester4	229	8.40	34.89
Ester4 Fotocopia	173	5.15	18.40
Ester5	229	9.19	36.43
Ester5 Fotocopia	172	4.35	14.55

Tabla 5.1: Resumen cuantitativo de las métricas calculadas en imágenes reales y fotocopias.

5.2.5.4 Análisis de los resultados

El estudio revela una serie de diferencias sistemáticas entre imágenes reales y sus correspondientes fotocopias:

- **Textura (varianza del Laplaciano):** en todos los casos analizados, las fotocopias han presentado una textura inferior a la de sus equivalentes reales. En casos extremos como `ester3` o `ester5`, la reducción es superior al 50%, lo que indica una pérdida significativa de detalle estructural. Este comportamiento está alineado con la hipótesis de que las superficies impresas, al ser planas, carecen de la complejidad tridimensional que caracteriza a un rostro real.
- **Variación de profundidad:** aunque en algunos pares la diferencia absoluta ha sido baja (por ejemplo, `pepe1`), en muchos otros la fotocopia ha mostrado un rango claramente inferior (`ester3`: 146 frente a 227). Esto apunta a una compresión o aplanamiento de la geometría facial en el proceso de impresión.
- **Desviación típica de profundidad:** esta métrica ha reflejado con consistencia una menor dispersión en las copias impresas, reforzando la conclusión anterior desde un enfoque estadístico. Los casos `ester3`, `ester4` y `ester5` destacan especialmente por su baja variabilidad, indicando uniformidad en la superficie que podría facilitar su detección.
- **Casos limítrofes:** en pares como `pepe1` o `pepe2`, las diferencias han sido sutiles, lo que sugiere que una clasificación binaria basada en una única métrica no sería fiable. Sin embargo, cuando se analizan en conjunto, los tres indicadores ofrecen una distinción más robusta entre ambas clases.

5.2.5.5 Conclusiones y líneas futuras

Los resultados obtenidos han permitido evidenciar patrones repetitivos que diferencian cuantitativamente a las imágenes reales de las fotocopias:

- Las imágenes reales presentan sistemáticamente mayor textura, mayor rango de profundidad y mayor dispersión.
- Las fotocopias tienden a tener superficies planas, menos detalladas y con una distribución de profundidad más uniforme.
- Las gráficas refuerzan visualmente estas diferencias y pueden utilizarse como base para el desarrollo de clasificadores basados en umbrales múltiples o técnicas de aprendizaje automático.

Este análisis establece un punto de partida para una clasificación supervisada, donde estas métricas pueden ser empleadas como atributos en modelos como árboles de decisión, SVM o redes neuronales. No obstante, para consolidar esta dirección, será necesario recopilar un conjunto de datos más amplio y diverso, además de establecer mecanismos de normalización y validación cruzada de los resultados.

5.3 Reflexión Final sobre los Resultados

A lo largo de este capítulo se han presentado y analizado los resultados obtenidos en las dos grandes líneas experimentales del proyecto: el uso de redes generativas antagónicas (GANs) y el análisis de profundidad mediante estimación monocular con *MiDaS*. Ambos enfoques han proporcionado evidencias empíricas que contribuyen a responder a los objetivos planteados, si bien desde perspectivas y niveles de madurez distintos.

Sobre el sistema GAN

El sistema basado en GANs ha demostrado capacidad para generar imágenes con estructura facial coherente, especialmente en resoluciones medias y altas. Se ha confirmado que el generador puede aprender progresivamente representaciones visuales realistas, y que el discriminador es sensible a estas mejoras, como lo refleja la evolución de su precisión.

Sin embargo, también se han observado limitaciones relevantes: el sistema es vulnerable al sobreajuste del discriminador, y muestra cierta inestabilidad en la pérdida adversarial. Aunque no se ha evidenciado un colapso de modo severo, se ha detectado una pérdida de diversidad en algunas ejecuciones prolongadas. En este sentido, el sistema se ha comportado como una herramienta útil para experimentar con generación facial, pero aún no se encuentra listo para ser integrado como módulo robusto de detección.

Sobre el análisis de profundidad con MiDaS

El análisis de profundidad monocular, por el contrario, ha mostrado un mayor grado de solidez y aplicabilidad inmediata. Las métricas extraídas sobre las ROIs faciales (variación de

profundidad, textura mediante el Laplaciano y desviación típica) han permitido diferenciar con claridad muchas de las imágenes reales de sus respectivas fotocopias. Estas diferencias han sido consistentes y han podido visualizarse tanto en gráficas agregadas como en histogramas individuales.

Uno de los hallazgos más relevantes ha sido que las fotocopias tienden a presentar superficies más planas y uniformes en los mapas de profundidad, lo cual se refleja en menores niveles de textura y dispersión. Este comportamiento es coherente con la naturaleza bidimensional de la imagen impresa y ha permitido establecer una base cuantitativa sólida para futuros sistemas de clasificación.

Cabe destacar que la validez del análisis se ha mantenido bajo condiciones controladas de interior, donde las variaciones de fondo e iluminación son limitadas. Se ha descartado el uso de imágenes exteriores para evitar introducir ruido y ambigüedad en la interpretación de los resultados.

Síntesis comparativa de ambos enfoques

Mientras que las GANs han sido exploradas desde una óptica generativa, enfocadas en la creación de muestras realistas y en la comprensión del comportamiento adversarial, el análisis de profundidad se ha dirigido más directamente a la caracterización y detección de ataques por presentación. En este sentido, MiDaS ha ofrecido resultados más inmediatos y prácticos, aunque también más dependientes del entorno de captura y la calidad de segmentación de la ROI.

Ambas líneas han aportado aprendizajes complementarios: por un lado, el conocimiento técnico necesario para entrenar modelos generativos controlados y evaluar su estabilidad; por otro, la utilidad de las métricas de profundidad como posibles atributos discriminativos en tareas de clasificación.

Perspectiva integrada

Los resultados obtenidos permiten concluir que la combinación de técnicas de análisis estructural, como las basadas en profundidad, con mecanismos adversariales —por ejemplo, empleando GANs para generar ejemplos de ataque o para reforzar el entrenamiento del sistema de detección— podría ser una línea futura prometedora.

La siguiente sección de discusión se encargará de examinar en mayor profundidad las implicaciones técnicas y prácticas de estos hallazgos, así como las posibles estrategias para superar las limitaciones detectadas. Esta reflexión final busca, por tanto, cerrar el capítulo de resultados ofreciendo una visión comparativa y crítica, sin anticipar las conclusiones definitivas ni propuestas concretas de mejora que se abordarán posteriormente.

6 Discusión

6.1 Interpretación General de los Resultados

6.1.1 Análisis de Resultados en el Contexto del Proyecto

A lo largo del trabajo se han desarrollado dos enfoques distintos para abordar la detección de ataques por suplantación facial: uno basado en redes generativas antagónicas (GANs) y otro en el análisis de mapas de profundidad mediante MiDaS. Cada uno ha ofrecido resultados útiles en su contexto y ha permitido validar parcialmente las hipótesis de partida.

En la línea de GANs, se ha conseguido entrenar modelos capaces de generar imágenes sintéticas de rostros con un grado razonable de realismo, especialmente a partir de la época 400. La calidad visual ha mejorado de forma progresiva, y el generador ha aprendido a reproducir estructuras faciales coherentes. El discriminador, por su parte, ha alcanzado altos niveles de precisión al inicio del entrenamiento, aunque con tendencia al sobreajuste en fases avanzadas, como se ha evidenciado en los tests cruzados con datos no vistos. En algunas ejecuciones también se ha detectado pérdida de diversidad en las muestras generadas, asociada a un inicio de colapso de modo.

El análisis de profundidad ha ofrecido una aproximación complementaria, centrada en la extracción de características estructurales a partir de una sola imagen. Mediante MiDaS, se han generado mapas de profundidad que han permitido comparar imágenes reales y fotocopiadas. En todos los casos analizados, las imágenes auténticas han mostrado mayor variabilidad en los valores de profundidad, más textura y una dispersión estadística más alta en la región facial. Estas diferencias se han confirmado visualmente y mediante métricas objetivas como la desviación típica o la varianza del Laplaciano.

En conjunto, ambos enfoques han aportado perspectivas diferentes pero complementarias sobre el problema. Las GANs han servido para modelar el comportamiento de generadores y discriminadores en contextos simulados, mientras que el análisis estructural ha ofrecido herramientas para detectar suplantaciones en datos reales.

6.1.2 Relación con Problemas de Verificación Facial

Los resultados obtenidos se han vinculado de forma directa con los retos actuales en verificación facial. El uso de GANs ha permitido generar ejemplos de posible ataque que simulan escenarios de fraude mediante imágenes sintéticas. Esto ha facilitado estudiar el comportamiento del sistema frente a entradas manipuladas, y ha puesto de manifiesto la necesidad de controlar el equilibrio entre generador y discriminador para evitar situaciones de sobreentrenamiento o falta de generalización.

Por otro lado, el análisis de mapas de profundidad ha permitido detectar patrones comunes en imágenes impresas, como superficies planas o pérdida de detalle estructural. Estas propiedades han sido explotadas para diferenciar visualmente imágenes reales de suplantaciones, y

podrían integrarse como módulos de verificación adicionales en sistemas biométricos.

Ambos enfoques han contribuido a reforzar la detección de ataques de presentación desde dos ángulos distintos: la generación sintética y el análisis estructural. Esta dualidad abre la puerta a sistemas más robustos, que combinen la capacidad de simular amenazas con la posibilidad de identificarlas mediante propiedades físicas de la imagen.

6.2 Limitaciones y Desafíos Encontrados

A lo largo del desarrollo del proyecto se han identificado múltiples obstáculos que han influido tanto en el diseño experimental como en la interpretación de los resultados. Estas limitaciones han abarcado desde aspectos técnicos hasta restricciones relacionadas con los datos y el entorno de ejecución. Aunque algunas ya se han mencionado en la Sección 4.6, esta sección las recoge y contextualiza desde una perspectiva más interpretativa, vinculándolas directamente con el rendimiento observado en las distintas fases experimentales.

6.2.1 Dificultades en el Entrenamiento GAN

El entrenamiento de redes generativas antagónicas ha supuesto uno de los mayores retos del proyecto. A pesar de haber logrado generar imágenes sintéticas con estructura facial coherente, se han experimentado varias dificultades relacionadas con la estabilidad y la convergencia del modelo.

En primer lugar, se ha observado una fuerte sensibilidad a la inicialización aleatoria y a los hiperparámetros, especialmente la tasa de aprendizaje y el tamaño del vector latente. Pequeñas variaciones en estos valores han derivado en comportamientos divergentes, afectando la calidad de las imágenes generadas o provocando oscilaciones en la pérdida adversarial.

También se ha detectado una tendencia al sobreajuste por parte del discriminador en etapas avanzadas del entrenamiento. Aunque al inicio ha alcanzado niveles de precisión elevados, su rendimiento sobre datos de test ha disminuido progresivamente, tal como se ha reflejado en las ejecuciones de `gan4.py`. Esta pérdida de capacidad de generalización ha limitado la utilidad práctica del discriminador como componente de verificación robusto.

Además, en algunas ejecuciones prolongadas se ha identificado un inicio de colapso de modo, manifestado en una pérdida de diversidad de las imágenes generadas. Este fenómeno ha comprometido la capacidad del generador para representar la variedad real del conjunto de datos, afectando negativamente alrealismo de las muestras producidas.

Por último, desde el punto de vista operativo, se ha requerido una monitorización constante del entrenamiento. La combinación de pérdida no estable, dependencia del preprocesamiento y necesidad de ajustes manuales ha dificultado la automatización completa del proceso.

6.2.2 Restricciones del Análisis de Profundidad

En cuanto al análisis basado en mapas de profundidad con MiDaS, se han encontrado limitaciones técnicas y metodológicas relevantes. Si bien el enfoque ha demostrado ser útil para distinguir entre imágenes reales y fotocopiadas, su rendimiento ha dependido en gran medida de la calidad visual de las imágenes de entrada.

Concretamente, se han producido errores de segmentación facial en casos con iluminación deficiente, desenfoques o encuadres inadecuados. Estos fallos han dificultado la extracción

precisa de la región de (ROI) y, en consecuencia, han afectado a la fiabilidad de las métricas de profundidad y textura calculadas sobre ella.

También se ha observado que en ciertos pares de imágenes el modelo MiDaS ha generado mapas de profundidad excesivamente planos, especialmente en escenas de baja complejidad visual o en copias impresas con poco contraste. Este comportamiento ha limitado la sensibilidad del análisis en esos casos.

Por otro lado, los umbrales empleados para evaluar la variación de profundidad y la textura han tenido un carácter exploratorio. Aunque han servido para identificar diferencias significativas en la mayoría de ejemplos, no han sido universalmente válidos para todos los sujetos y condiciones de captura, lo que ha restringido la posibilidad de establecer una regla de decisión definitiva.

Adicionalmente, el conjunto de imágenes utilizado ha sido de tamaño reducido y ha sido capturado íntegramente por el propio autor. Esta circunstancia, aunque ha permitido un control riguroso sobre las condiciones de prueba, limita la representatividad del análisis y dificulta la generalización de los resultados a escenarios reales más amplios. La ausencia de muestras externas, con mayor diversidad de sujetos, dispositivos y entornos, reduce la validez externa del sistema y subraya la necesidad de evaluar su rendimiento con datos independientes.

Estas restricciones han puesto de manifiesto la necesidad de desarrollar mecanismos adaptativos de segmentación y análisis, así como de ampliar la base de datos con imágenes más diversas y realistas que reflejen mejor las condiciones operativas de un sistema biométrico real.

6.3 Robustez, Fiabilidad y Generalización

Esta sección aborda una reflexión crítica sobre la solidez de los modelos desarrollados, su comportamiento ante datos no vistos y su potencial aplicabilidad en contextos reales. Se ha evaluado la robustez de las GANs frente a variaciones internas del entrenamiento, así como la capacidad de generalización del análisis de profundidad con MiDaS ante situaciones de suplantación.

6.3.1 Robustez del Sistema GAN

A lo largo de las distintas etapas experimentales, el sistema GAN ha mostrado un comportamiento razonablemente estable, especialmente en términos de convergencia visual y evolución progresiva del generador. Se ha comprobado que el entrenamiento ha producido imágenes cada vez más detalladas a lo largo de las épocas, manteniéndose la diversidad entre muestras y sin haber evidenciado colapso de modo severo en las ejecuciones más relevantes.

La robustez estructural del modelo se ha manifestado en la repetibilidad de los resultados cualitativos a través de múltiples ejecuciones. A pesar de la inicialización aleatoria y del impacto inherente del azar en cada entrenamiento, se ha observado una evolución consistente del generador, con mejoras claras en la definición facial y la expresión emocional en el caso del conjunto FER2013. No obstante, se han detectado oscilaciones notables en la precisión del discriminador, lo cual ha puesto de manifiesto la sensibilidad del modelo a pequeños cambios en la configuración o en el conjunto de datos.

Además, se ha constatado que la estabilidad del entrenamiento ha estado condicionada por la resolución de entrada y el volumen de datos disponibles. En resoluciones más altas, la arquitectura ha requerido ajustes adicionales para mantener un equilibrio entre la calidad visual y la precisión discriminativa. Esta dependencia ha evidenciado la necesidad de calibrar cuidadosamente los hiperparámetros y ha limitado en parte la aplicabilidad directa del modelo a escenarios más complejos sin un ajuste previo.

En resumen, el sistema GAN ha demostrado una robustez aceptable dentro del marco experimental, pero también ha evidenciado vulnerabilidades ante cambios externos, especialmente en fases avanzadas del entrenamiento, donde la pérdida adversarial ha tendido a volverse más inestable y el discriminador ha mostrado signos de sobreajuste.

6.3.2 Capacidad de Generalización del Modelo de Profundidad

La evaluación de la capacidad de generalización del modelo de análisis de profundidad se ha centrado en determinar si las métricas estructurales extraídas han permitido distinguir de forma coherente entre imágenes reales y fotocopias más allá de los ejemplos inicialmente observados. Dado que el modelo MiDaS ha sido preentrenado sobre conjuntos de datos diversos, ha partido con una capacidad intrínseca de extrapolación a escenas desconocidas. Sin embargo, su rendimiento específico sobre imágenes faciales ha dependido en gran medida de las condiciones particulares de captura.

Durante el estudio se ha comprobado que el modelo ha sido capaz de generar mapas de profundidad útiles en una mayoría significativa de los casos. En particular, cuando las imágenes han sido tomadas en interiores y con una calidad visual adecuada, los histogramas y métricas sobre la región facial han revelado diferencias claras y repetitivas entre las imágenes reales y las impresas. Este comportamiento sugiere una respuesta generalizable del sistema bajo condiciones similares a las del entorno experimental.

Ahora bien, este alcance se ha visto restringido por dos factores clave. Por un lado, el conjunto de imágenes utilizado ha sido limitado en tamaño y en variabilidad, al haber sido generado manualmente por el propio autor. Por otro lado, la sensibilidad del modelo frente a condiciones desfavorables —como iluminación deficiente, desenfoque o encuadres mal ajustados— ha provocado que, en determinados casos, los mapas de profundidad no hayan reflejado una estructura discernible, dificultando el análisis posterior.

Por tanto, aunque se ha verificado una capacidad de generalización funcional en escenarios controlados, no puede garantizarse que este comportamiento se mantenga ante imágenes más heterogéneas o capturadas en contextos operativos reales. Se concluye que el modelo, en su estado actual, ofrece un punto de partida prometedor para tareas de verificación, pero su aplicación práctica requerirá tanto una base de datos más diversa como mejoras en los mecanismos de preprocesamiento y segmentación, que aseguren una mayor robustez frente a la variabilidad de entrada.

6.4 Implicaciones Metodológicas y Técnicas

6.4.1 Lecciones Aprendidas sobre Aprendizaje Adversarial

El desarrollo y evaluación de modelos basados en aprendizaje adversarial han permitido identificar una serie de implicaciones metodológicas de relevancia para futuros trabajos en

generación y discriminación de imágenes faciales. A lo largo de las distintas etapas experimentales se ha comprobado que las GANs presentan un comportamiento altamente sensible a la configuración inicial, tanto en lo relativo a la arquitectura como a los hiperparámetros de entrenamiento.

Se ha aprendido que la estabilidad durante el entrenamiento no puede darse por sentada y que requiere una monitorización continua, tanto a nivel numérico (mediante métricas de pérdida y precisión) como cualitativo (mediante análisis visual de las imágenes generadas). En particular, se ha verificado que el generador ha necesitado un número considerable de épocas para alcanzar una síntesis visualmente coherente, y que el discriminador, en muchas ocasiones, ha mostrado una evolución oscilante, tendente al sobreajuste en fases avanzadas.

Otra lección destacada ha sido la importancia de la diversidad de datos en el proceso de entrenamiento. En aquellos casos en los que el conjunto de entrenamiento ha sido escaso o poco representativo, el modelo ha mostrado una pérdida de generalidad, incluso con arquitecturas teóricamente adecuadas. Además, se ha comprobado que el uso de imágenes intermedias durante el entrenamiento ha facilitado la detección temprana de colapso de modo y ha servido como mecanismo indirecto de validación del progreso del modelo.

En conjunto, se ha evidenciado que el enfoque adversarial, aunque potente y flexible, requiere una disciplina experimental rigurosa y una capacidad de diagnóstico constante. Su aplicabilidad a entornos reales depende no solo de la calidad de las imágenes generadas, sino también de la robustez de los discriminadores y de su resistencia al sobreajuste. Estas conclusiones refuerzan la necesidad de incorporar prácticas como la parada temprana, la validación cruzada y el análisis visual sistemático en cualquier implementación seria de modelos GAN en tareas de seguridad biométrica.

6.4.2 Valor del Análisis Estructural Basado en Profundidad

La aplicación del análisis estructural mediante mapas de profundidad ha aportado una dimensión complementaria al enfoque generativo, y ha permitido explorar con éxito una vía alternativa para la detección de intentos de suplantación facial. A través del modelo MiDaS y de las técnicas asociadas de segmentación y extracción de métricas, se ha podido evaluar de forma efectiva la estructura tridimensional de los rostros, incluso a partir de imágenes RGB convencionales.

Se ha aprendido que este tipo de análisis permite detectar patrones distintivos entre imágenes auténticas y copias impresas, en particular gracias a la menor variabilidad de profundidad y textura presente en estas últimas. Esta propiedad ha ofrecido un criterio cuantificable de evaluación, complementario al juicio visual, y ha demostrado ser útil para discriminar entre clases sin necesidad de entrenamiento supervisado directo sobre ejemplos falsificados.

Asimismo, se ha comprobado que la fiabilidad del análisis ha dependido en gran medida de la calidad de las máscaras faciales y del correcto aislamiento de la región de interés. Esto ha resaltado el valor de contar con procesos de segmentación robustos, especialmente cuando se pretende aplicar estas técnicas a escenarios operativos no controlados.

Desde el punto de vista técnico, el uso de un modelo preentrenado como MiDaS ha permitido acelerar la validación experimental, evitando la necesidad de entrenar desde cero y facilitando la integración del sistema en flujos de trabajo existentes. Sin embargo, también se ha puesto de manifiesto que la calidad del análisis puede verse afectada por factores como la

iluminación, la resolución de entrada o la calidad de la impresión, lo que requiere considerar estos elementos en el diseño de futuras soluciones.

En síntesis, el análisis de profundidad ha demostrado ser una herramienta valiosa para enriquecer la evaluación estructural de imágenes biométricas. Su integración con otros enfoques, como los adversariales, puede constituir una línea metodológica sólida para el desarrollo de sistemas de detección de fraudes más robustos, especialmente en contextos donde la variabilidad del entorno representa un reto adicional.

6.5 Síntesis de Aportes

6.5.1 Valor Científico y Técnico del Trabajo

A lo largo de este trabajo se ha llevado a cabo una exploración rigurosa de dos enfoques complementarios para abordar el problema de la verificación facial frente a ataques por presentación: las redes generativas antagónicas (GANs) y el análisis estructural mediante mapas de profundidad. Esta doble aproximación ha permitido no solo validar técnicas consolidadas en contextos específicos, sino también evaluar de forma crítica sus limitaciones y puntos fuertes en un entorno biométrico concreto.

Desde una perspectiva científica, se ha aportado un marco experimental reproducible que ha permitido observar y documentar fenómenos clave del aprendizaje adversarial, como el colapso de modo, el sobreajuste del discriminador y la oscilación de las métricas de entrenamiento. Se ha evidenciado cómo estos fenómenos se manifiestan en distintas configuraciones y con diferentes conjuntos de datos, lo cual ha servido para ilustrar el comportamiento práctico de las GANs más allá de los resultados idealizados presentes en la literatura.

En el plano técnico, se ha implementado un sistema funcional de generación de rostros sintéticos a partir de ruido, así como un discriminador entrenado para distinguir imágenes reales de generadas. El entrenamiento ha sido complementado con una estrategia de visualización periódica, almacenamiento de métricas y generación de resultados intermedios, lo que ha facilitado el seguimiento preciso de la evolución de los modelos.

Además, se ha diseñado e integrado un sistema completo de análisis de profundidad que, mediante el uso del modelo MiDaS, ha permitido generar mapas tridimensionales relativos y extraer métricas cuantificables sobre la estructura facial. Este sistema ha sido aplicado con éxito a pares de imágenes reales e impresas, y ha permitido identificar patrones repetitivos de aplanamiento y pérdida de textura en las copias, sentando así las bases para su posible uso como herramienta de detección automática.

Ambas líneas de trabajo han sido ejecutadas con independencia metodológica, pero con una vocación de integración a futuro, lo que demuestra la madurez técnica alcanzada en el desarrollo. En conjunto, se ha ofrecido una solución coherente, documentada y escalable para la evaluación de ataques por suplantación facial, contribuyendo con ello al avance práctico y experimental en este campo.

6.5.2 Relevancia dentro del Campo de la Verificación Facial

El presente trabajo se ha situado en el contexto de los sistemas de verificación facial, un ámbito de creciente importancia en aplicaciones de seguridad, control de acceso y autenticación biométrica. Dentro de este marco, los ataques por presentación —como el uso de

fotocopias o imágenes generadas— representan una amenaza real y cada vez más sofisticada, frente a la cual los sistemas tradicionales muestran vulnerabilidades significativas.

En este sentido, la investigación desarrollada ha tenido una clara orientación aplicada, ya que ha abordado no solo la generación de ejemplos sintéticos, sino también su análisis estructural como vía para fortalecer la detección de fraudes. El uso de GANs ha permitido estudiar cómo un sistema puede enfrentarse a imágenes generadas artificialmente, mientras que la estimación de profundidad ha proporcionado un recurso adicional para detectar carencias físicas en imágenes impresas.

Se ha demostrado que, aunque los sistemas analizados aún requieren mejoras para su despliegue en entornos operativos, han ofrecido un punto de partida válido y prometedor. La propuesta metodológica ha puesto de relieve la utilidad de incorporar técnicas complementarias—visuales, métricas y estructurales—en la verificación facial, especialmente en escenarios donde la mera información RGB puede ser insuficiente.

La relevancia de este trabajo reside también en su valor pedagógico y experimental: ha servido como base para documentar y entender con detalle el comportamiento de técnicas avanzadas en un problema realista, contribuyendo al conocimiento aplicado sobre ciberseguridad biométrica y generando evidencia útil para futuras investigaciones, desarrollos tecnológicos o integraciones con sistemas híbridos de defensa contra ataques por presentación.

7 Conclusiones

Este capítulo recoge los hallazgos más relevantes derivados del desarrollo del presente Trabajo de Fin de Máster, centrado en el estudio de técnicas avanzadas para la detección de ataques por suplantación facial mediante redes generativas antagónicas (GANs) y análisis estructural basado en mapas de profundidad. A modo de cierre, se revisan los objetivos inicialmente planteados, se sintetizan las principales contribuciones técnicas y metodológicas, y se valora críticamente el alcance de los resultados obtenidos.

Además, se exponen las limitaciones más significativas detectadas a lo largo del proceso experimental, tanto en la línea generativa como en la analítica, y se reflexiona sobre el impacto potencial de estas técnicas en entornos reales de verificación biométrica. Esta revisión final no solo permite contrastar el grado de cumplimiento de los objetivos propuestos, sino también establecer una base argumental para justificar las propuestas de mejora recogidas en el capítulo siguiente.

7.1 Cumplimiento de Objetivos

A lo largo del desarrollo de este trabajo se han perseguido una serie de objetivos orientados a investigar soluciones computacionales ante ataques por suplantación facial. El objetivo general ha consistido en evaluar la viabilidad de dos enfoques complementarios —las redes generativas antagónicas (GANs) y el análisis estructural mediante mapas de profundidad— como mecanismos de detección. A continuación, se expone el grado de cumplimiento alcanzado en cada uno de los objetivos específicos formulados inicialmente:

- **Desarrollar un modelo GAN capaz de generar rostros sintéticos con cierto nivel de realismo.** Este objetivo se ha cumplido de forma razonable. Se ha conseguido que el generador aprenda a producir rostros con estructuras faciales coherentes a partir de ciertas épocas de entrenamiento (en especial a partir de la época 400). No obstante, la calidad visual alcanzada no ha igualado la de modelos del estado del arte, y en algunas ejecuciones se han observado síntomas de colapso de modo y pérdida de diversidad.
- **Entrenar un discriminador que permita distinguir entre imágenes reales y sintéticas.** Este objetivo se ha alcanzado parcialmente. Aunque el discriminador ha demostrado una alta capacidad de diferenciación en fases iniciales del entrenamiento, su rendimiento se ha deteriorado en fases avanzadas, mostrando sobreajuste y menor eficacia ante datos no vistos. Esto ha limitado su aplicabilidad como detector robusto en contextos reales.
- **Aplicar técnicas de análisis de profundidad para comparar imágenes reales y falsificadas.** Este objetivo se ha abordado pero no puede considerarse plenamente

cumplido. Se ha implementado un análisis estructural basado en mapas de profundidad generados con MiDaS y se han extraído métricas relevantes. Sin embargo, la baja diversidad del conjunto de imágenes y la sensibilidad del modelo a condiciones como la iluminación o el encuadre han condicionado la fiabilidad del análisis. En varios casos, los mapas generados han sido excesivamente planos o poco representativos, lo que ha dificultado la extracción de conclusiones sólidas y generalizables.

- **Evaluar la robustez y la capacidad de generalización de ambos enfoques.** Este objetivo se ha cumplido de forma limitada. Se han realizado pruebas con ejemplos no vistos dentro del conjunto reducido disponible, pero no se ha contado con una base de datos suficientemente heterogénea como para validar adecuadamente la generalización de los sistemas. Tanto las GANs como el modelo de profundidad han mostrado un comportamiento razonable en condiciones controladas, pero su estabilidad ante variabilidad real no ha podido comprobarse de forma concluyente.
- **Identificar fortalezas, debilidades y posibles líneas de mejora en los sistemas evaluados.** Este objetivo sí se ha cumplido con amplitud. A lo largo del análisis y discusión se han documentado detalladamente los principales retos técnicos encontrados (como el sobreajuste, la inestabilidad en el entrenamiento adversarial o las limitaciones en la segmentación facial) y se han propuesto líneas de mejora concretas, como la necesidad de ampliar la base de datos, optimizar los procesos de preprocesamiento o integrar enfoques híbridos.

En definitiva, aunque los sistemas desarrollados no han alcanzado un nivel de madurez suficiente para su despliegue en entornos operativos, sí han permitido avanzar de forma significativa en la exploración de enfoques alternativos para la detección de ataques por suplantación. El trabajo ha generado resultados reproducibles, lecciones técnicas valiosas y una base experimental sobre la que construir futuras mejoras. En ese sentido, el TFM ha cumplido con su propósito formativo y ha contribuido, en su escala, al campo de la verificación facial biométrica.

7.2 Principales Contribuciones

A lo largo del desarrollo de este trabajo se han realizado varias aportaciones técnicas y metodológicas que, si bien se han ejecutado en un entorno experimental controlado, han resultado relevantes tanto desde el punto de vista formativo como aplicable. Las principales contribuciones se resumen a continuación:

- **Implementación de un sistema GAN para generación de rostros sintéticos.** Se ha diseñado y entrenado un modelo generativo basado en redes adversariales que ha sido capaz de producir imágenes faciales con una estructura coherente. A pesar de sus limitaciones en cuanto a diversidad y calidad visual, el modelo ha demostrado ser funcional, y ha permitido observar fenómenos propios de estos sistemas, como el colapso de modo o el sobreajuste del discriminador. Esta implementación ha aportado una plataforma básica reutilizable para estudios futuros de generación facial o simulación de ataques por presentación.

- **Desarrollo de un discriminador adversarial entrenado sobre imágenes reales y sintéticas.** Aunque su capacidad de generalización ha sido limitada, se ha construido un discriminador que ha alcanzado altos niveles de precisión durante fases iniciales del entrenamiento. Su comportamiento ha ofrecido información relevante sobre la dinámica del entrenamiento adversarial y ha permitido analizar el impacto de los datos y la configuración sobre el sobreajuste.
- **Aplicación del modelo MiDaS para análisis estructural de imágenes faciales.** Se ha integrado con éxito un sistema de análisis basado en profundidad, que ha permitido generar mapas tridimensionales relativos sobre imágenes RGB. A partir de estos mapas, se han extraído métricas objetivas (como la desviación típica y la varianza del Laplaciano) que han reflejado diferencias estructurales entre imágenes reales y copias impresas. Esta contribución ha demostrado el potencial del análisis de profundidad como técnica no invasiva para detectar ataques por suplantación.
- **Diseño de una metodología de evaluación comparada entre enfoques generativos y estructurales.** Se ha llevado a cabo un análisis sistemático de dos enfoques conceptualmente distintos para abordar la detección de fraudes en verificación facial. Este contraste ha permitido identificar sus puntos fuertes y debilidades, así como establecer criterios preliminares para su posible combinación en soluciones híbridas.
- **Generación de un conjunto de datos anotado para pruebas exploratorias.** Aunque limitado en volumen y diversidad, se ha creado un pequeño dataset de imágenes faciales reales e impresas, capturado bajo condiciones controladas y documentado con su correspondiente estructura de pares. Este recurso, junto con el código implementado, constituye una base sobre la que otros experimentos pueden construirse o reproducirse.

Estas contribuciones, aunque acotadas en alcance por la naturaleza del trabajo, han servido para profundizar en el conocimiento práctico de técnicas actuales en visión por computador y biometría, y han sentado las bases para su posterior mejora, evaluación o integración en soluciones más robustas.

7.3 Impacto de los Resultados

Los resultados obtenidos a lo largo de este trabajo han ofrecido una base empírica útil para reflexionar sobre el potencial de ciertos enfoques en la detección de ataques por presentación en sistemas de verificación facial. Aunque el alcance experimental ha estado limitado por la escala del proyecto y la disponibilidad de datos, se han derivado implicaciones relevantes tanto a nivel práctico como metodológico.

En primer lugar, el uso de redes generativas antagónicas ha permitido simular escenarios de fraude mediante la creación de imágenes faciales sintéticas. Esto ha proporcionado un entorno controlado para estudiar el comportamiento de discriminadores entrenados frente a manipulaciones visuales. Si bien los modelos no han alcanzado un nivel de realismo comparable al de sistemas del estado del arte, se ha evidenciado su utilidad como herramienta pedagógica y de simulación para futuras pruebas de robustez en sistemas biométricos.

Por su parte, el análisis basado en mapas de profundidad ha mostrado un mayor potencial de aplicabilidad inmediata. Se ha demostrado que es posible extraer indicadores estructurales

que permitan diferenciar imágenes reales de impresas, incluso en ausencia de entrenamiento específico sobre ejemplos falsificados. Esta capacidad de generalizar ciertas propiedades físicas del rostro ha supuesto un avance metodológico con aplicaciones potenciales en escenarios de bajo coste, donde el uso de sensores adicionales no sea viable.

Además, se ha puesto de manifiesto la viabilidad de aplicar herramientas preentrenadas, como MiDaS, en contextos específicos como la ciberseguridad biométrica. Aunque su rendimiento ha dependido fuertemente de la calidad de las imágenes de entrada, se ha verificado que, bajo condiciones adecuadas, puede contribuir de forma efectiva a la evaluación no invasiva de autenticidad.

En conjunto, los resultados del trabajo han evidenciado que enfoques complementarios —como los aquí explorados— pueden servir como puntos de partida para diseñar soluciones híbridas más robustas frente a intentos de suplantación. Las técnicas desarrolladas no son directamente desplegables en entornos productivos, pero sí han ofrecido una base técnica transferible que puede integrarse, adaptarse y mejorarse en el marco de investigaciones o desarrollos más avanzados.

Por tanto, el impacto principal del trabajo reside en haber construido una aproximación exploratoria con potencial de extensión, y en haber generado evidencia útil para guiar decisiones futuras sobre qué técnicas merecen ser escaladas o combinadas en sistemas biométricos reales.

7.4 Limitaciones Detectadas

Durante el desarrollo del trabajo se han identificado diversas limitaciones que han condicionado el alcance de los resultados y la generalización de las conclusiones. Estas limitaciones, ya analizadas en profundidad en capítulos anteriores, se resumen aquí desde una perspectiva crítica.

En primer lugar, el análisis estructural basado en profundidad ha estado limitado por la escasa diversidad del conjunto de imágenes utilizadas, ya que todas ellas han sido capturadas manualmente en el marco del proyecto. Aunque esta decisión ha permitido un mayor control sobre las condiciones de captura, también ha restringido la representatividad del conjunto y, por tanto, la capacidad de extrapolar los resultados a escenarios operativos más diversos. La falta de variabilidad en sujetos, dispositivos y condiciones ha afectado especialmente a la evaluación de la generalización del sistema basado en MiDaS.

En cambio, en la línea de redes generativas antagónicas se han empleado conjuntos de datos públicos más variados, como FER2013 o Conjunto de Celebridades Anotadas a Gran Escala (Large-scale CelebFaces Attributes Dataset) (CelebA)¹ Liu y cols. (2015), lo que ha permitido explorar una mayor riqueza de expresiones y configuraciones faciales. No obstante, incluso en este caso se han encontrado limitaciones relacionadas con la estabilidad del entrenamiento, la sensibilidad a los hiperparámetros y fenómenos como el colapso de modo. Además, el discriminador ha mostrado una marcada tendencia al sobreajuste, con pérdida de rendimiento frente a ejemplos no vistos.

¹CelebA (Large-scale CelebFaces Attributes Dataset) es un conjunto de datos ampliamente utilizado en tareas de reconocimiento facial y aprendizaje profundo. Contiene más de 200,000 imágenes de celebridades anotadas con atributos faciales, poses y marcas clave.

Otra limitación relevante ha sido la dependencia de condiciones visuales óptimas para obtener resultados fiables en el análisis de profundidad. Factores como la iluminación deficiente, el desenfoque o un encuadre inadecuado han provocado que en varios casos los mapas generados fuesen excesivamente planos o poco informativos. Esto ha afectado la consistencia de las métricas extraídas y ha dificultado establecer umbrales de decisión robustos.

Finalmente, cabe señalar que, al tratarse de un estudio con una orientación exploratoria, no se ha llevado a cabo una evaluación cuantitativa exhaustiva mediante métricas estandarizadas como la tasa de aciertos, precisión o

glsauc². Esta ausencia ha limitado la comparabilidad de los resultados frente a métodos del estado del arte y refuerza la necesidad de realizar validaciones más sistemáticas en trabajos futuros.

En conjunto, estas limitaciones han reflejado tanto los desafíos técnicos como las restricciones metodológicas y de disponibilidad de recursos, y marcan un punto de partida claro para futuras líneas de mejora.

7.5 Cierre y Perspectiva

La realización de este trabajo ha permitido adquirir una comprensión sólida de dos enfoques relevantes para la detección de ataques por suplantación facial: el uso de redes generativas antagónicas (GANs) y el análisis estructural basado en mapas de profundidad. A lo largo del proceso, se han enfrentado numerosos retos técnicos y metodológicos que han exigido una constante revisión de decisiones, así como una actitud crítica hacia los resultados obtenidos.

Desde el punto de vista práctico, se ha logrado implementar un sistema de generación y discriminación de imágenes faciales funcional, así como un módulo de análisis de profundidad capaz de detectar diferencias estructurales entre imágenes reales y falsificadas en escenarios controlados. No obstante, el sistema en su estado actual debe considerarse una primera aproximación exploratoria, más orientada a validar la viabilidad conceptual que a ofrecer una solución directamente aplicable en contextos operativos reales.

El desarrollo del trabajo ha puesto de relieve la importancia de factores como la calidad y diversidad de los datos, la estabilidad del entrenamiento adversarial o la robustez de los mecanismos de segmentación. Asimismo, ha evidenciado el valor de combinar distintas perspectivas técnicas —sintética y estructural— para abordar un problema de seguridad biométrica desde ángulos complementarios.

A nivel personal y formativo, la experiencia ha supuesto un ejercicio riguroso de integración de conocimientos previos, adaptación a tecnologías complejas y toma de decisiones fundamentadas. El aprendizaje derivado del trabajo abarca tanto aspectos técnicos específicos como competencias transversales en diseño experimental, análisis crítico y documentación científica.

De cara al futuro, el trabajo deja abiertas múltiples líneas de mejora y extensión que se detallan en el siguiente capítulo. Entre ellas, destacan la necesidad de ampliar la base de datos con muestras más representativas, optimizar los procesos de entrenamiento y evaluación, e investigar estrategias híbridas que integren ambos enfoques de forma coordinada. Estas

²AUC (Área Bajo la Curva) es una medida que indica qué tan bien un modelo puede distinguir entre clases. Un valor alto significa mejor capacidad para diferenciar correctamente entre positivos y negativos.

perspectivas constituyen una base sólida para futuras investigaciones que busquen avanzar en la detección robusta de fraudes biométricos en sistemas de verificación facial.

8 Trabajos Futuros

Este capítulo recoge diversas propuestas de mejora, ampliación y nuevas líneas de investigación que podrían derivarse del trabajo desarrollado. Estas propuestas se fundamentan tanto en las limitaciones detectadas durante el desarrollo como en oportunidades identificadas para incrementar la robustez, generalización y aplicabilidad del sistema propuesto.

8.1 Ampliación y Validación Experimental

Una línea clara de evolución consiste en ampliar la validación experimental del sistema. En este trabajo se ha utilizado un conjunto de datos controlado, lo cual facilita el entrenamiento inicial y la evaluación de la viabilidad del enfoque. Sin embargo, para poder valorar adecuadamente su capacidad de generalización, es necesario probar el sistema con bases de datos públicas y más diversas como Labeled Faces in the Wild (LFW) Li (2020), CASIA-FASD Liu y cols. (2015) o CASIA-FASD¹ Zhang y cols. (2012), que ofrecen una mayor variabilidad en términos de identidades, condiciones de iluminación, poses y expresiones faciales.

Asimismo, resulta de interés evaluar el rendimiento frente a ataques más sofisticados, como impresiones de alta resolución, vídeos pregrabados o máscaras 3D, así como en condiciones de captura menos controladas. Estas pruebas permitirían validar la robustez del sistema en entornos reales. Por último, sería relevante incluir una comparación con métodos del estado del arte en benchmarks reconocidos, utilizando métricas estandarizadas para valorar el desempeño relativo del sistema.

8.2 Mejoras del Sistema GAN

8.2.1 Ajuste de Arquitectura y Hiperparámetros

Otra línea de mejora importante reside en la propia arquitectura de la GAN empleada. Resulta interesante explorar variantes más avanzadas como Red Generativa Adversarial Convolucional Profunda (Deep Convolutional GAN) (DCGAN), Red Generativa Adversarial Estilizada (Style-based GAN) (StyleGAN) o Wasserstein GAN con Penalización de Gradiente (Wasserstein GAN with Gradient Penalty) (WGAN-GP)², que han demostrado una mayor capacidad para generar imágenes de alta calidad y estabilidad durante el entrenamiento.

¹CASIA-FASD (Face Anti-Spoofing Database) es una base de datos pública ampliamente utilizada en la investigación de detección de ataques de presentación, que incluye vídeos reales y falsificados en distintos tipos de ataque como impresiones, reproducciones en pantalla y máscaras.

²DCGAN (Deep Convolutional GAN), stylegan y WGAN-GP (Wasserstein GAN with Gradient Penalty) son variantes avanzadas de redes generativas adversariales que han mejorado la calidad y estabilidad del entrenamiento. StyleGAN, por ejemplo, es conocido por generar rostros sintéticos con alto realismo, mientras que WGAN-GP mejora la convergencia mediante una nueva función de pérdida.

Además, se podrían probar diferentes técnicas de normalización, funciones de activación o estrategias de regularización que optimicen el rendimiento del modelo.

Complementariamente, la realización de una búsqueda sistemática de hiperparámetros —mediante métodos como *grid search*, *random search* o algoritmos evolutivos— podría permitir alcanzar un rendimiento más óptimo y reproducible, mejorando la calidad de las imágenes generadas y la eficacia del clasificador.

8.2.2 Uso del Generador como Fuente de Datos

Una posibilidad adicional consiste en utilizar el generador entrenado como una fuente de datos sintéticos. Estas imágenes podrían emplearse para enriquecer los conjuntos de entrenamiento de clasificadores tradicionales o basados en aprendizaje profundo, mejorando su capacidad para detectar ataques de presentación. De este modo, se podría aumentar la diversidad del conjunto de datos sin necesidad de adquirir más ejemplos reales, lo cual resulta especialmente valioso en dominios donde la obtención de datos está restringida.

Asimismo, podrían generarse ejemplos específicos de ataques simulados, contribuyendo así al desarrollo de sistemas más robustos frente a diferentes tipos de amenazas.

8.3 Fortalecimiento del Análisis de Profundidad

8.3.1 Evaluación con Datos Más Diversos

Una de las principales limitaciones del análisis estructural mediante mapas de profundidad ha sido la escasa diversidad del conjunto de imágenes utilizado. Para mejorar la validez externa del sistema, sería fundamental ampliar la base de datos con imágenes capturadas en condiciones más variadas. Esto incluiría incorporar sujetos con diferente morfología facial, imágenes obtenidas con distintos dispositivos (como webcams, cámaras móviles o sistemas de vigilancia), y entornos de iluminación diversos.

Adicionalmente, validar el sistema con datos de terceros —provenientes de conjuntos públicos o colaboraciones externas— permitiría contrastar su rendimiento en escenarios no controlados. Esta evaluación contribuiría a identificar posibles sesgos, comprobar la generalización del enfoque y consolidar su aplicabilidad en contextos reales.

8.3.2 Automatización y Optimización

Otra línea de mejora relevante consiste en automatizar el proceso de análisis. Actualmente, la detección de la región de interés (ROI) facial requiere intervención manual en algunos casos, lo cual limita la escalabilidad del sistema. Implementar una detección automática, robusta y fiable de la ROI mediante modelos de segmentación o detección facial modernos (por ejemplo, Red Neuronal Convolutacional en Cascada Multitarea (Multi-task Cascaded Convolutional Neural Network) (MTCNN)³ o MediaPipe) podría resolver este problema.

³MTCNN (Multi-task Cascaded Convolutional Neural Network) es un detector facial basado en redes neuronales profundas que permite localizar con precisión rostros en imágenes, incluso en condiciones de iluminación o ángulo desfavorables.

Asimismo, se propone adaptar de forma dinámica los umbrales empleados para calcular las métricas de profundidad y textura, ajustándolos según las condiciones de entrada. Esto permitiría mantener una sensibilidad adecuada ante variaciones de calidad, escala o resolución, y podría mejorar la fiabilidad del sistema sin requerir un ajuste fijo para cada escenario.

8.4 Despliegue en Entornos Reales

8.4.1 Funcionamiento en Tiempo Real

Con vistas a una aplicación práctica, sería interesante adaptar los modelos desarrollados para su funcionamiento en tiempo real. Para ello, habría que evaluar su latencia y eficiencia computacional en dispositivos reales, como cámaras de vigilancia, terminales de acceso o sistemas portátiles.

Además, se podría explorar la optimización de los modelos mediante técnicas como la cuantización, la poda o la conversión a formatos ligeros (por ejemplo, TensorFlow Lite y Formato Abierto de Intercambio de Redes Neuronales (Open Neural Network Exchange) (ONNX))⁴. Esto permitiría reducir los requisitos de hardware y facilitar su integración en sistemas con recursos limitados, sin comprometer significativamente la precisión del análisis.

8.4.2 Desarrollo de Herramienta Interactiva

Una extensión práctica del trabajo consistiría en desarrollar una herramienta gráfica interactiva que permita visualizar los resultados de análisis, ejecutar pruebas sobre nuevas imágenes o simular ataques de presentación. Esta herramienta podría servir como plataforma de demostración, facilitar la validación del sistema por parte de evaluadores o ser empleada en contextos educativos.

El diseño de una interfaz amigable permitiría también que usuarios no técnicos —como miembros de un jurado, personal de seguridad o responsables de validación— pudieran interpretar los resultados de forma intuitiva, lo cual ampliaría el alcance del sistema más allá del ámbito puramente técnico.

8.5 Consideraciones Éticas y de Seguridad

El uso de modelos generativos en contextos biométricos plantea desafíos éticos y de seguridad que no deben ser ignorados. Uno de los riesgos más evidentes es el mal uso del generador para la creación de contenidos falsificados, como los llamados *deepfakes*. Aunque en este trabajo el generador se ha empleado con fines experimentales, la tecnología subyacente puede ser explotada con fines maliciosos si no se establecen mecanismos de control.

Por ello, sería conveniente estudiar medidas que permitan auditar o restringir el uso de los modelos generativos, tales como la introducción de marcas de agua digitales, el registro de las ejecuciones o la verificación de la autenticidad del contenido generado. También sería pertinente abrir un debate sobre la responsabilidad ética y legal asociada al desarrollo y

⁴TensorFlow Lite y ONNX son formatos optimizados para ejecutar modelos de aprendizaje profundo en dispositivos con recursos limitados, como móviles, cámaras o sistemas embebidos.

distribución de este tipo de herramientas, especialmente cuando se integran en sistemas de identificación o verificación biométrica.

Reflexión Final

Las líneas propuestas en este capítulo representan una continuación lógica y necesaria del trabajo realizado. Abordan tanto las limitaciones técnicas detectadas como las oportunidades metodológicas que han surgido a lo largo del proceso experimental. Aunque los sistemas desarrollados han demostrado un potencial prometedor, su consolidación requiere aún un esfuerzo adicional orientado a la validación empírica, la optimización de procesos y la consideración de aspectos éticos.

Avanzar en estas direcciones permitiría no solo reforzar la robustez y aplicabilidad de las soluciones propuestas, sino también contribuir de manera más sólida al desarrollo de tecnologías biométricas seguras, responsables y adaptables a los desafíos emergentes en ciberseguridad facial.

Bibliografía

- Abderrahmane Nitaj, A. N. (2023). Applications of neural network-based ai in cryptography. *MDPI Cryptography*. doi: <https://doi.org/10.3390/cryptography7030039>
- Akamai. (2024). *Modelo de seguridad zero trust*. Descargado de <https://www.akamai.com/es/glossary/what-is-zero-trust>
- Akhtar, N., y Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6, 14410–14430. doi: 10.1109/ACCESS.2018.2807385
- Alabdullah, A. M., Qahwaji, D., y Khan, I. U. (2016). Fraud detection in online financial transactions using machine learning algorithms. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 7(7), 1–7. doi: 10.14569/IJACSA.2016.070701
- Alshamrani, M. I., y Alkasassbeh, G. M. (2021). Machine learning in cybersecurity: A review. *IEEE Access*, 9, 121296–121312. doi: 10.1109/ACCESS.2021.3109635
- Arpita Gupta, P. T. D. N., Galgotias College. (2020). *Cryptography using artificial intelligence*. Descargado de https://www.researchgate.net/publication/354834635_Cryptography_Using_Artificial_Intelligence
- ArticAI. (2025). *La ciberseguridad, un pilar crítico en la era de la inteligencia artificial*. Descargado de <https://www.articai.es/evolucion-tecnologia-ia-ciberseguridad-2025/> (Consultado el 12 de febrero de 2025)
- Ayerbe, A. (2020). La ciberseguridad y su relación con la inteligencia artificial. *Real Instituto Elcano*. Descargado de <https://www.realinstitutoelcano.org/analisis/la-ciberseguridad-y-su-relacion-con-la-inteligencia-artificial/>
- Birkl, R., Wofk, D., y Müller, M. (2023). Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*. Descargado de <https://arxiv.org/abs/2307.14460>
- Carlini, N., y Wagner, D. (2017). Towards evaluating the robustness of neural networks. En *Ieee symposium on security and privacy (sp)* (pp. 39–57). doi: 10.1109/SP.2017.49
- Choithani, T., Chowdhury, A., Patel, S., Patel, P., Patel, D., y Shah, M. (2024). A comprehensive study of artificial intelligence and cybersecurity on bitcoin, cryptocurrency, and banking system. *Springer AI Journal*. doi: 10.1007/s40745-022-00433-5
- Cibersafety. (2024). *Guía de recuperación ante ransomware*. Descargado de <https://cibersafety.com/recuperacion-ransomware-pasos-empresa/> (Consultado el 12 de febrero de 2025)

- CISA. (2025). *Malware, phishing, and ransomware*. Descargado de <https://www.cisa.gov/topics/cyber-threats-and-advisories/malware-phishing-and-ransomware> (Consultado el 12 de febrero de 2025)
- Cloudflare. (2024). *¿qué es la defensa en profundidad? / seguridad en capas*. Descargado de <https://www.cloudflare.com/es-es/learning/security/glossary/what-is-defense-in-depth/>
- Cohen, G. R. J. B. O. A. W. R. M. A. S. R. (2023). A survey on explainable artificial intelligence for cybersecurity. *IEEE Transactions on Artificial Intelligence*. doi: 10.1109/TNSM.2023.3282740
- Comillas, U. P. (2024). *Confidencialidad, integridad y disponibilidad*. Descargado de <https://ciberseguridad.comillas.edu/confidentiality-integrity-and-availability/>
- Contributors, W. (2025). *Michael i. jordan*. https://en.wikipedia.org/wiki/Michael_I._Jordan. Descargado de https://en.wikipedia.org/wiki/Michael_I._Jordan (Última actualización: 28 de enero de 2025)
- Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., y Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1), 53–65. doi: 10.1109/MSP.2017.2765202
- DataSunrise. (2024). *Ejemplos de confidencialidad, integridad y disponibilidad*. Descargado de <https://www.datasunrise.com/es/centro-de-conocimiento/ejemplos-de-confidencialidad-integridad-disponibilidad/>
- Directores de Seguridad. (2024). *Biometría y seguridad: principales vulnerabilidades*. Descargado de <https://directoresdeseguridad.es/2024/12/18/biometria-y-seguridad-principales-vulnerabilidades/> (Consultado el 12 de febrero de 2025)
- Elastic. (2024). *¿qué es ciberseguridad?* Descargado de <https://www.elastic.co/es/what-is/cybersecurity> (Consultado el 12 de febrero de 2025)
- ENAE International Business School. (2024). *La importancia de la ciberseguridad en el mundo digital*. Descargado de <https://www.ename.es/blog/la-importancia-de-la-ciberseguridad-en-el-mundo-digital> (Consultado el 12 de febrero de 2025)
- First, O. (2024). *Zero trust security architecture: A blueprint for modern cybersecurity approach*. Descargado de <https://objectfirst.com/es/guides/data-security/zero-trust-security-architecture/>
- Fortinet. (2024). *Tríada cia: confidencialidad, integridad y disponibilidad*. Descargado de <https://www.fortinet.com/lat/resources/cyberglossary/cia-triad>
- Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., y Greenspan, H. (2018). Synthetic data generation using gan for improved liver lesion classification. *IEEE Transactions on Medical Imaging*, 38(3), 915–925. doi: 10.1109/TMI.2018.2863040

- Generative adversarial networks: Inteligencia artificial & ciberseguridad (1 de 2). (2018). *El Lado del Mal*. Descargado de <https://www.elladodelmal.com/2018/11/generative-adversarial-networks.html>
- González, J. (2020). La inteligencia artificial y la robótica: sus dilemas sociales, éticos y jurídicos. *Revista Mexicana de Derecho*, 7(3), 49-72. Descargado de https://www.scielo.org.mx/scielo.php?pid=S2448-51362020000300049&script=sci_arttext
- Goodfellow, I., y cols. (2013). *Fer2013 - facial expression recognition challenge*. <https://www.kaggle.com/datasets/msambare/fer2013>. (Accedido en mayo de 2025)
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. En *Advances in neural information processing systems (neurips)* (pp. 2672–2680). Curran Associates, Inc. Descargado de <https://papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>
- Goodfellow, I. J., Shlens, J., y Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*. Descargado de <https://arxiv.org/abs/1412.6572>
- Grewal, B. S. P. G. J. K. (2023). Advances and challenges in cryptography using artificial intelligence. *IEEE Transactions on Cryptography*. doi: 10.1109/I2CT57861.2023.10126338
- Greydon, T. (2022). Detecting deepfakes: An overview of techniques and challenges. *ACM Computing Surveys*, 55(4), 1–39. doi: 10.1145/3495249
- Hong, M., Xia, Y., Yu, S., y Wang, H. (2022). A review of generative adversarial networks in security and privacy applications. *IEEE Access*, 10, 48671–48691. doi: 10.1109/ACCESS.2022.3169607
- IBM. (2024). *¿qué es la ciberseguridad?* Descargado de <https://www.ibm.com/es-es/topics/cybersecurity> (Consultado el 12 de febrero de 2025)
- Insights, I. (2024). *Evaluaciones completas de riesgos de ia: Fortalecimiento de la ciberseguridad y cumplimiento para el éxito empresarial.* Descargado de <https://insights.integrity360.com/es/comprehensive-ai-risk-assessments-enhancing-cyber-security-compliance-for-business-success>
- Javaid, A., Niyaz, Q., Sun, W., y Alam, M. (2016). A deep learning approach for network intrusion detection system. *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies*, 21–26. doi: 10.4108/eai.3-12-2015.2262516
- Jonathan Blackledge, N. M. (2020). *Applications of artificial intelligence to cryptography*. doi: <https://doi.org/10.14738/tmlai.83.8219>
- Kaggle. (2024). *Kaggle: Your machine learning and data science community*. Descargado de <https://www.kaggle.com/> (Accedido el 13 de mayo de 2025)

- Li, J. (2020). *Labeled faces in the wild (lfw) dataset*. <https://www.kaggle.com/datasets/jessicali9530/lfw-dataset>. (Accedido el 15 de mayo de 2025)
- Liu, Z., Luo, P., Wang, X., y Tang, X. (2015). *Deep learning face attributes in the wild*. <https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>. (Accessed: 2025-05-15)
- Los usos criminales de los 'deepfakes' se disparan: estafas, pornografía y suplantación de identidad. (2024). *El País*. Descargado de <https://elpais.com/proyecto-tendencias/2024-10-02/los-usos-criminales-de-los-deepfakes-se-disparan-estafas-pornografia-y-suplantacion-de-identidad.html>
- Lysenko, N. K. K. V. O. T. Y., Serhii; Bobro. (2024). The role of artificial intelligence in cybersecurity: Automation of protection and detection of threats. *ProQuest Journal of AI Security*. doi: 10.46852/0424-2513.1.2024.6
- Meghna Manoj Nair, A. K. T., Atharva Deshmukh. (2024). Artificial intelligence for cyber security: Current trends and future challenges. En *Cybersecurity trends*. Wiley. doi: 10.1002/9781394213948
- Moosavi-Dezfooli, S.-M., Fawzi, A., y Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2574–2582. doi: 10.1109/CVPR.2016.282
- Naser, J. Y. I. A.-Z. S. B. A. S. S. M. A. (2023). A survey of cryptographic algorithms with deep learning. *AIP Advances*. doi: 10.1063/5.0133509
- Núñez, J. F. (2024). *Marco de ciberseguridad nist 2.0: Guía integral para implementación*. Descargado de <https://es.linkedin.com/pulse/marco-de-ciberseguridad-nist-20-gu%C3%A3DA-integral-para-fern%C3%A1ndez-n%C3%BA%C3%B1ez-ewy9c>
- Otto, C. (2024). Así usan los ciberdelincuentes los 'deepfakes': de estafas bancarias a suplantaciones de voz y rostro. *El Confidencial*. Descargado de https://www.elconfidencial.com/tecnologia/innovacion/2024-01-15/deepfakes-ciberdelincuencia-estafas-banco_3780212/
- Pan, X., You, S., Yang, X., Wang, F., y Qian, C. (2021). A survey of generative adversarial networks and their applications in computer vision. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4797–4817. doi: 10.1109/TNNLS.2020.3011401
- País, E. (2024). *Reglamento de ia: una oportunidad para innovar con seguridad*. Descargado de <https://elpais.com/economia/negocios/2024-09-01/reglamento-de-ia-una-oportunidad-para-innovar-con-seguridad.html>
- Pirani. (2024). *Ciberseguridad: qué es, cómo funciona y su importancia*. Descargado de <https://www.piranirisk.com/es/academia/especiales/ciberseguridad-que-es-como-funciona-y-su-importancia> (Consultado el 12 de febrero de 2025)
- PricewaterhouseCoopers. (2021). *Guía de buenas prácticas en el uso de la inteligencia artificial ética*. Descargado de <https://www.pwc.es/es/publicaciones/tecnologia/assets/guia-buenas-practicas-uso-inteligencia-artificial-pwc-odiseia.pdf>

- Ramanpreet Kaur, T. K., Dušan Gabrijelčič. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Future Generation Computer Systems*. doi: <https://doi.org/10.1016/j.inffus.2023.101804>
- Ranftl, R., Bochkovskiy, A., y Koltun, V. (2020). *Midas: Real-time depth estimation*. <https://github.com/intel-is1/MiDaS>. (Accessed: 2025-05-13)
- Riesgos de ia y ciberseguridad. (2025). *Malwarebytes*. Descargado de <https://www.malwarebytes.com/es/cybersecurity/basics/risks-of-ai-in-cyber-security>
- Russell, S., y Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Sakurai, I. M. S. D. H. T. K. (2021). Neural networks-based cryptography: A survey. *IEEE Transactions on Neural Networks*. doi: 10.1109/ACCESS.2021.3109635
- Singh, S. (2024). Enhancing cyber security using quantum computing and artificial intelligence: A review. *ResearchGate Review*. doi: 10.48175/IJARSCT-18902
- Sommer, R., y Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *IEEE Symposium on Security and Privacy*, 305–316. doi: 10.1109/SP.2010.25
- Sutton, R. S., y Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
- Zhang, Z. L., Yan, J., y Li, S. Z. (2012). Casia face anti-spoofing database. En *Ieee conference on computer vision and pattern recognition workshops (cvprw)* (pp. 1–7). Descargado de <http://www.cbsr.ia.ac.cn/english/FaceAntiSpoofDatabases.asp> (Accessed: 2025-05-15)

Lista de Acrónimos y Abreviaturas

AES	Estándar de Encriptación Avanzado (Advanced Encryption Standard).
ANN	Redes Neuronales Artificiales (Artificial Neural Networks).
AUC	Área bajo la Curva (Area Under the Curve).
CelebA	Conjunto de Celebridades Anotadas a Gran Escala (Large-scale CelebFaces Attributes Dataset).
CIA	Confidencialidad, Integridad y Disponibilidad (Confidentiality, Integrity and Availability).
CNN	Red Neuronal Convolucional (Convolutional Neural Network).
DCGAN	Red Generativa Adversarial Convolucional Profunda (Deep Convolutional GAN).
DDoS	Ataque Distribuido de Denegación de Servicio (Distributed Denial of Service).
DPT	Transformador de Predicción Densa (Dense Prediction Transformer).
EC	Computación Evolutiva (Evolutionary Computation).
FER2013	Conjunto de Datos de Reconocimiento de Emociones Faciales (Facial Expression Recognition 2013).
FGSM	Método de Signo de Gradiente Rápido (Fast Gradient Sign Method).
FP32	Punto Flotante de 32 Bits (32-bit Floating Point).
GAN	Redes Adversariales Generativas (Generative Adversarial Networks).
GDPR	Reglamento General de Protección de Datos.
GPU	Unidad de Procesamiento Gráfico (Graphics Processing Unit).
GRU	Unidad Recurrente Gated (Gated Recurrent Unit).
IA	Inteligencia Artificial.
IDS	Sistema de Detección de Intrusiones (Intrusion Detection System).

IEEE	Institute of Electrical and Electronics Engineers.
IoT	Internet de las Cosas (Internet of Things).
IPS	Sistema de Prevención de Intrusiones (Intrusion Prevention System).
LFW	Labeled Faces in the Wild.
LSTM	Memoria a Largo y Corto Plazo (Long Short-Term Memory).
LWE	Aprendizaje con Errores (Learning With Errors).
MiDaS	Estimación de Profundidad Monocular (Monocular Depth Estimation).
ML	Aprendizaje Automático (Machine Learning).
MNIST	Instituto Nacional de Estándares y Tecnología Modificado (Modified National Institute of Standards and Technology).
MTCNN	Red Neuronal Convolucional en Cascada Multitarea (Multi-task Cascaded Convolutional Neural Network).
NIST	Instituto Nacional de Estándares y Tecnología (National Institute of Standards and Technology).
ONNX	Formato Abierto de Intercambio de Redes Neuronales (Open Neural Network Exchange).
QKD	Distribución de Claves Cuánticas (Quantum Key Distribution).
ReLU	Unidad Lineal Rectificada (Rectified Linear Unit).
RFID	Identificación por Radiofrecuencia (Radio Frequency Identification).
RGB	Rojo Verde Azul (Red Green Blue).
RNN	Red Neuronal Recurrente (Recurrent Neural Network).
ROI	Región de Interés (Region of Interest).
RSA	Rivest-Shamir-Adleman.
SCP	Protocolo Seguro de Copia (Secure Copy Protocol).
SGSI	Sistema de Gestión de Seguridad de la Información.
SIEM	Gestión de Información y Eventos de Seguridad (Security Information and Event Management).
SSH	Protocolo Seguro de Conexión Remota (Secure Shell).
StyleGAN	Red Generativa Adversarial Estilizada (Style-based GAN).
SVM	Máquina de Vectores de Soporte (Support Vector Machine).

TFG	Trabajo Final de Grado.
TFM	Trabajo Final de Máster.
WGAN	Wasserstein Generative Adversarial Network.
WGAN-GP	Wasserstein GAN con Penalización de Gradiente (Wasserstein GAN with Gradient Penalty).
XAI	Inteligencia Artificial Explicable (Explainable Artificial Intelligence).