Análisis Forense de Imágenes y Vídeos Manipulados (Deepfake)

Trabajo de Fin de Máster en Ciberseguridad

U.N.E.D.

Realizado por: Azucena Martín Castaño

Dirigido por: María de los Llanos Tobarra Abad

Codirigido por: Antonio Robles Gómez

A mi padre

Agradecimientos

En primer lugar, quiero dar las gracias a mi tutora Mª de los Llanos Tobarra, por darme la confianza que necesitaba para enfrentarme a este reto, en el que me ha guiado, corregido y asesorado y por su trato cercano y correcto. Sin duda este trabajo le debe una parte muy importante. Y al profesor Antonio Robles Gómez, quien me ha dado el último empujón necesario para finalizar.

Agradecer también el apoyo, ánimo y confianza a toda mi familia y en especial a mi pareja por ser el sufridor más directo, quien se ha visto vetado de vacaciones y salidas.

Resumen

Vivimos en una era donde el tiempo es oro. La sociedad ha creado la necesidad de estar informada de cualquier acontecimiento al segundo en el que se produce un evento Además, se espera, sobre todo dentro del ámbito de las redes sociales, que un líder político o personaje público influyente, se manifieste y de su opinión o versión sobre lo que está pasando o sobre lo que piensa al respecto. Esta urgencia, es la que lleva la mayoría de los consumidores, a dar por válidas las imágenes que están visualizando en sus pantallas. Olvidando el hecho de que, entre otras cosas, el abaratamiento de la tecnología y la Inteligencia Artificial, conllevan a que cada vez es más sencillo falsificar un vídeo o una imagen. En los últimos años han proliferado aplicaciones, cada vez más versátiles y aptas para un público no profesional, capaces de falsificar un vídeo con bastante calidad si se le dedica un tiempo al aprendizaje. Este hecho, no pasa inadvertido para aquellos que, por diversos motivos, quiere extender un bulo. Así pues, cada vez es más necesario contar con tecnología y métodos que ayuden en la labor de verificación de las imágenes. Este trabajo pretende realizar un repaso sobre las manipulaciones más frecuentes y presentar las principales técnicas de comprobación.

Palabras claves

Deepfake, Redes sociales, Compresión JPEG, Redes neuronales convolucionales (CNN, Convolultion Neutral Network), MesoNet

Abstract

We live an era where the time is a gold. Society has created the need to be informed when any event has happened at the same moment. Besides, it is expected, especially on social medias, that a politician leader or influential person gives his/her opinion about the matter. This urgency is what most consumer to accept the images which they are viewing on their screens are true. Forgetting the fact, among other things, the decrease in the price of technology and the Artificial Intelligent involve forging a video or image is getting easier. The last year, versatile applications have proliferated, and they are suitable for unprofessional people. People capable to fake a video with high quality if they spend a little time to learning. This fact does not go unnoticed by some who pretend to extend a hoax for his /her own reason. Thus, it is more necessary to have technology and methods which help us to verify an image. This work pretends to do a review over more usual tampering and show some verification techniques.

Keywords

Deepfake, social media, JPEG compression, Convolutional neutral network (CNN), MesoNet.

Índice

Índice de figuras

Índice de tablas

Acrónimos

- **API** American Press Institute
- **APP1** Application Marker Segment 1
- **BTT** Back Propagation Through Time
- **CANet** Compressión Artifact Network
- **CNN** Convolutian Neural Network
- **DCT** Discrete Cosene Transform
- **DFT** Discrete Fournier Transform
- **EOI** End Of Imagen
- **EXIF** Exchangeable Image File
- **FPS** Frame Per Second
- **GAN** Generarive Adversary Network
- **GOP** Groups Of Imagen
- **IA** Artificial Intelligente
- **JPEG** Joint Photographic Experts Group
- **LSTM** Long-Short Term Memory
- **MCNet** Maniupulation Classification Network
- **R** Recall
- **RNN** Recurrent Neural Networks
- **SOI** Start Of Imagen
- **SOS** Strart Of Stream
- **SVM** Support Vector Machines
- **VANet** Visual Artifact Network

1. Introducción

1.1 Motivación

Para entender las razones detrás del auge de las Deepfakes, es necesario comprender el contexto en el que nos encontramos y estudiar las posibles principales causas de esta evolución.

Es evidente que hoy en día gran parte de la población dispone dispositivos móviles, equipados generalmente con cámaras de altas prestaciones y gran capacidad de almacenamiento que permiten que la calidad del vídeo y de la imagen sea muy alta, a unos precios asequibles. Según [1] el 76,7% de la población utiliza el móvil para consumir contenidos digitales y un 64,7 % usa el ordenador. Como segundo factor importante, podemos destacar las redes sociales. Las cuales juegan un papel cada vez más importante en nuestra sociedad. La población no solo las utiliza para estar en contacto con sus allegados, sino que cada vez se consume más información a través de las mismas [3]. Un ejemplo claro se muestra en [2], el cual indica que uno de cada dos adolescentes hace uso de las redes sociales para encontrar la información. Y por último reseñar que, gracias al uso de aplicaciones de Inteligencia Artificial (IA, *Intelligent Artificial),* el retoque fotográfico o la creación de vídeos manipulados, está al alcance de cualquiera, sin ser profesional, obteniendo muy buenos resultados.

El término de deepfake se utilizó por primera vez en 2017. Este fenómeno obtuvo su nombre de un usuario anónimo de la plataforma Reddit, que se hizo llamar "deepfakes" y que compartió los primeros deepfakes colocando a celebridades desconocidas en un videoclip para adultos [4]. Deepfakes (un neologismo que combina los términos "aprendizaje profundo" y "falso") son videos (o imágenes) sintéticos en los que el rostro de una persona (y, opcionalmente, también la voz) se reemplaza por la imagen de otra persona utilizando tecnologías de aprendizaje profundo [5].

La facilidad de creación de deepfakes y la rapidez de distribución a través de las plataformas sociales, junto con la falta de verificación por parte de la mayoría de los usuarios, hacen que este tipo de información engañosa circule libremente por todas las redes sociales. El público quiere difundir su opinión, su información, no callar [6] amparados en la libertad de expresión, junto con la falta de regularización interna que las plataformas sociales tienen para que alguien suba o difunda una información falsa, hace que como apunta el autor en [7], la mayor amenaza no reside en el hecho de que el receptor de la información sea engañado, sino en que la información misma pierda toda la credibilidad.

1.2 DeepFake y usos malignos

Las plataformas de redes sociales nos permiten estar en contacto no sólo con nuestros familiares más cercanos. Permiten además de mantener ese contacto con personas que han pasado por nuestras vidas y ya no están cercas y que de otra manera no lo haríamos, gracias a su facilidad de contacto y rapidez. Pero, son precisamente estas dos características las que hacen que una deepfake se propague. La mayoría de los usuarios en las redes sociales, dan por válida una imagen o un vídeo manipulado solo por el hecho de que quien lo difunde es afín a sus aficiones, ideologías políticas o religiosas [8]. Estas circunstancias son explotadas por grupos de intereses que hacen uso de las redes (sobre todo Twitter) para atacar contra una minoría tanto étnica como de género [9]

En [10], el estudio muestra que la mayoría de deepfake de contenido pornográfico son protagonizados por mujeres mientras que los vídeos interpretados por hombres son referencias políticas o humorísticas. En la pasada campaña electoral a la presidencia en Estados Unidos de 2016, se alegó que las noticias falsas podrían haber sido fundamentales en la elección del presidente

Trump [11]. Existen bots[1] y trolls [2], ciberataques y piratas informáticos, cuentas secuestradas, propaganda y pretendientes de todo tipo. Abundan las mentiras y los mentirosos que influyen directamente en nuestra capacidad como ciudadanos para percibir, comprender y responder a la realidad [12].

Hay al menos cuatro tipos principales de deepfake productores: 1) comunidades de aficionados a los deepfake; 2) actores políticos como gobiernos extranjeros, y varios activistas; 3) otros actores malévolos como estafadores; y 4) actores legítimos, como las compañías de televisión [13].

Aunque es cierto que el uso de deepfake, en su primera aparición tiene una connotación negativa, no todos los usos son precisamente así. Por ejemplo, con esta tecnología la industria cinematográfica, video juegos o televisiva han tenido un gran avance en películas por ejemplo de animación y recientemente podemos ver en las pantallas de televisión el famoso anuncio de Lola Flores que caracteriza al mito, donde no sólo se ve la imagen de la famosa cantautora, sino que oímos su voz y su forma de hablar [14] .

1.3 Análisis forense en Media

Hay gran contenido digital sobre los personajes famosos en plataformas de streaming como Youtube, entrevistas, además de todo lo relacionado con su vida laboral, esta es la razón por la que un atacante podría manipular un vídeo o una foto en cualquier pose y sustituirla en otra, para que parezca que dice o hace una cosa distinta a la que sucedió en la imagen origen. Cualquier dato puede tener un valor significativo. Esto puede incluir los parámetros de captura (geolocalización, modelo de cámara, nivel de iluminación del entorno), post-procesamiento (calidad de compresión, filtros aplicados) o manipulación explícita. Por ejemplo, localizar modificaciones en una región de la imagen [15]. Cada vez es más necesaria la verificación de manipulación de deepfakes [16].

1 Programa informático para la realización de tareas repetitivas sin la intervención humana utilizados para atacar sistemas, envíos masivos de correos o vulnerar sitios webs. https://latam.kaspersky.com/resource-center/definitions/what-are-bots. Ultimo acceso 17/06/2022

2 Persona con identidad desconocida que buscan provocar de manera intencionada polémica y conflictos con la finalidad de divertirse, actuando en foros, blogs, o cualquier otro soporte online. https://metricool.com/es/trolls-redes-sociales/. Último acceso 17/06/2022.

Nuevos métodos de manipulación visual y de contenido aparecen cada día lo cual obliga a un continuo y constante desarrollo en las herramientas de análisis forense [17]. El análisis forense de imágenes se divide en dos grandes ramas: 1) autentificación de la imagen y 2) la identificación de las fuentes [18]. Según [19], aunque los algoritmos propuestos pueden dar una idea clara de identificación, todos presentan algún tipo de carencia que conlleva una identificación no tan clara del dispositivo. Por otro lado, en [21] se demuestra que, aumentado el mapa de características de la capa convolucional en redes neuronales, se añade una mejora significativamente esta tarea. Trabajo que facilita la identificación del dispositivo en el análisis forense, como evidencia en procedimientos legales o investigaciones criminales. No obstante, este trabajo no se centrará en la identificación del dispositivo, sino únicamente en la autentificación de la imagen.

1.4 Plan de trabajo

El presente trabajo se encuentra estructurado de la siguiente manera:

- **Capítulo 2.** Este capítulo está dividido en dos partes, la primera sección de planificación, en la que se exhiben los tiempos de elaboración del presente proyecto y las distintas fases en las que se ha realizado y en que han consistido cada una de las mismas, y una segunda sección de presupuesto, desglosado según la naturaleza del gasto y que correspondería a una simulación de los costes económicos que se hubieran imputado si este trabajo hubiera sido desarrollado por personal especializado en la materia.
- **Capítulo 3** Presentación del marco teórico de la captura, procesamiento y compresión de una imagen y vídeo, resaltando las más demandas por el mercado y los usuarios. Se resaltan las técnicas de manipulación más utilizas en según las bases teóricas expuestas y se introducen los esquemas de computación sobre los que fundamentas las metodologías de los distintos algoritmos que se exhiben en el mismo.
- **Capítulo 4.** Barrido de las técnicas más comunes utilizadas para la detección de imágenes falsificadas, que han sido manipulada alterando o eliminando alguna de las características propias de la imagen y que facilitan la tarea para los analistas forenses. Estos

procedimientos de detección se apoyan en los algoritmos de computación vistos en el capítulo previo.

- **Capítulo 5.** Exposición del experimento propio llevado a cabo, en el que se ponen en práctica los conocimientos teóricos adquiridos, y el cual plantea el uso de la arquitectura Mesonet, el conjunto de datos utilizados, junto con los resultados obtenidos.
- **Capítulo 6.** Conclusiones obtenidas tras el estudio y el experimento realizado y como contribuye éste a la comunidad dedicada al análisis forense y en su último epígrafe, trabajos futuros, en el que se definen las líneas de trabajo que no se han cubierto en el presente proyecto.

2. Planificación y presupuesto

Este capítulo se divide en dos secciones, la primera de planificación en la que se detalla el consumo de horas destinadas al proyecto y una segunda parte en la que especifica los costes económicos asociados a la elaboración en un entorno de trabajo.

2.1 Planificación

A continuación, se detallan las distintas fases en las que se ha dividido el trabajo para su realización. La estimación de tiempo del trabajo se inicia desde el momento que se realizó la petición de propuesta y finaliza en el momento que este trabajo es presentado para evaluación.

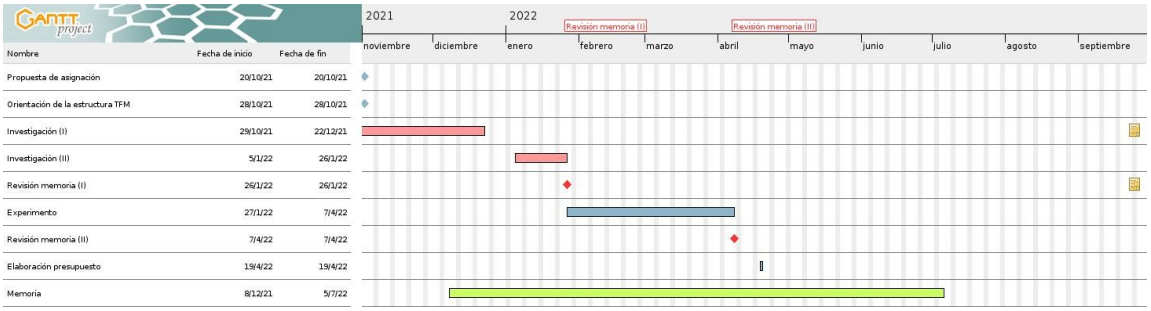

Figura 1: Gantt del proyecto y fases en las que se que se ha divido

La figura 1 superior, muestra el Gantt[3] con el ciclo de vida del proyecto en el que visualizan las diferentes fases. Como se puede apreciar, la memoria se creó al poco tiempo del inicio de investigación del trabajo y ha ido creciendo a lo largo del tiempo según se ha avanzado en conocimientos. Este crecimiento continuo conlleva a revisiones, las cuales son los hitos relevantes que se muestran en el Gantt. A continuación, se realiza una visión sobre las fases:

1. **Investigación**

   En esta fase se ha realizado una tarea de búsqueda de documentación, lectura y análisis de todos los artículos y publicaciones asociados a la materia del trabajo. No todos han sido incluidos en la bibliografía. Toda la documentación adquirida ha sido necesaria para obtener los conocimientos con el fin de llevar a cabo la ejecución del proyecto No solo ha aportado las enseñanzas a nivel técnico si no que ha ampliado la visión de la problemática con respecto a las deepfake. En el diagrama se aprecia la tarea segregada en dos partes y es debido al parón por el periodo navideño, así pues, la tarea se vuelve a retomar en enero del 2022. Esta tarea ha requerido sesiones de control por parte de la tutora, para enfocar el contenido y establecer unas bases.

   Durante el transcurso de esta tarea no se ha producido ningún incidente, puesto que existe una gran cantidad de fuentes con contenido público que se ha podido consultar sin problemas. La duración de la misma ha sido en total 55 días.

2. **Experimento**

   Este apartado presenta la fase práctica del trabajo. Contempla la configuración del espacio de trabajo y la demostración de los algoritmos planteados, junto con la preparación del conjunto de datos necesarios.

   En esta fase se han encontrado ciertos errores en la configuración del entorno por conflictos de librerías que no permitían una ejecución de los algoritmos correctamente o visualización de las imágenes y encontrar un dataset que fuera lo suficientemente variable para que la muestra fuera buena y se pudiera ejecutar en el entorno para que mostrase datos válidos con tiempos de ejecución razonable. Duración total de la tarea 51 días.

3 https://www.ganttproject.biz/. Ultimo acceso 15/06/2022

3. **Presupuesto**

   Esta tarea realiza una estimación de la carga económica que serían imputados en un entorno laboral. Duración 1 día.

4. **Memoria**

   Como se ha mencionado anteriormente, con el fin de organizar ideas, asimilar conocimientos, y empezar a estructurar toda la información recabada, se decide empezar a elaborar la memoria del proyecto al mes de comenzar la tarea de investigación, empezando con un esqueleto e ir creciendo en contenidos conforme se asientan conocimientos. Esta evolución hace necesaria la revisión, hitos marcados en el Gantt, por parte de la tutora con el objetivo de matizar y no divagar en el contenido.

2.2 Presupuesto

Para el cálculo de importes se establece que un mes consta de 4 semanas. El cómputo de horas final del proyecto son 300 horas. A continuación, se detalla el presupuesto del desarrollo. Dividido en 3 partidas:

- Presupuesto material intangible
- Presupuesto material tangible
- Presupuesto mano de obra

2.2.1 Presupuesto material intangible

Este aparado se compone de todas las licencias que se han pagado para la ejecución del proyecto y demás suministros intangibles para llevar a cabo el desarrollo, por ejemplo, gasto eléctrico, conexión a Internet, teléfono.

De acuerdo a las características dadas en el apartado 5., para el desarrollo se han utilizado las herramientas Python 3.8.10, keras 2.8.0 y Numpy 1.22.2. Todas las herramientas y librerías son públicas y de acceso gratuito, por lo que este coste no es imputable a los gastos del proyecto. Para la conexión a Internet se dispone de una contratación de fibra óptica con un operador nacional con una cuota establecida de 25€, por lo que el precio en la semana es de 6,25 €. En lo referente a la factura

de la luz se ha realizado una media del consumo de 15 semanas (300 horas) en el año anterior, añadiendo un porcentaje por el aumento del precio anual por lo que se establece un precio de 70€ al mes, que corresponde 17,5 € a la semana.. La tabla 1 muestra el resumen de los gastos.

Descripción	Cantidad (semanas)	Precio semana	Total
Software & librerías	0	0	0
Suministro eléctrico	15	17,5	262,5
Conexión Internet	15	6,25	93,75
		Total	**356,25**

Tabla 1: Presupuesto imputado a los costes intangibles

2.2.2 Presupuesto tangible

No se ha necesitado adquirir ningún equipo informático, ni cualquier otro equipo de trabajo, el trabajo se ha desarrollado con los medios ya disponibles, por lo tanto, no existe imputación de gastos en este apartado.

2.2.3 Presupuesto mano de obra

Este desglose contempla las horas de trabajo y el coste de las jornadas de trabajo imputadas para el resultado final del trabajo.

La duración total del proyecto se establece en 300 horas que equivalen a 3 meses y 3 semanas de trabajo (15 semanas). Se establece la jornada laboral de lunes a viernes, en horario de 17:00 a 21:00 horas, por lo que se trabajan 4 horas al día, que genera una carga de trabajo de 20 horas a la semana. El trabajo de búsqueda de documentación y transcripción de la parte teórica de la memoria puede ser desarrollado por un técnico auxiliar. La parte de experimento y análisis de las conclusiones puede ser llevada a cabo por un experto analista forense que finalice la documentación. De este modo la división de horas y carga de trabajo con los costes se pueden en la tabla 2.

Descripción	Nº. Semanas	Total horas	Precio hora	Total
Técnica	9	180	35	6300
Analista	6	120	100	12000
			Total	**18300**

Tabla 2: Presupuesto imputado a la mano de obra

2.2.4 Presupuesto final

A continuación, en la tabla 3, se muestra el presupuesto total del proyecto.

Descripción	Cantidad	Precio unitario	Total
Material intangible	1	356,25	356,25
Material tangible	0	0	0
Mano de obra	1	18300	18300
		Total	**18656,25**

Tabla 3: Presupuesto con los importes totales y finales

3. Estudio teórico

3.1 Imágenes digitales

Las redes sociales y numerosas páginas web suelen, acompañar su noticia o la información que van a compartir o cualquier tipo de exposición junto a una imagen para dar veracidad o sentido a la publicación en sí. En el análisis forense es importante identificar si esa imagen es falsa o no, pero para ello hay que conocer como es esa imagen y su composición e identificación, para a posteriori entrar en la estimación de si es o no real.

Una imagen es una matriz bidimensional definida por la función f*(x,y),* donde *(x,y)* son las coordenadas espaciales y *f* una función en cualquier par de coordenadas *(x,y)* es la intensidad de dicho punto [23]. Cada elemento de esa matriz representará la intensidad de la luz a la ubicación del sensor correspondiente. La figura 2 muestra el proceso de conversión de la captura de una imagen, las diferentes transformaciones que sufre para formar la imagen digital en este caso, se trata de una imagen a color. La luz penetra por la lente, los filtros especiales que separan la luz en rojo, verde y azul, en este caso la capa Bayer, pasa a los conversores analógicos-digital para así obtener la imagen digital.

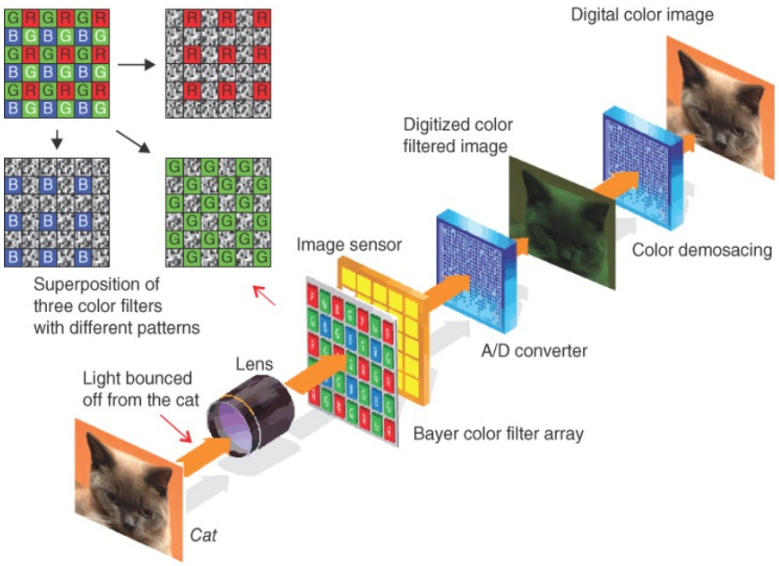

Figura 2: Esquema captura de una imagen por una cámara digital [22]

3.1.1 JPEG

La diversidad de formatos de los que podemos disponer para guardar las imágenes es amplio JPEG, Gif, PNG, tif/tiff, RAW, BMP. Sin embargo, este trabajo se centrará únicamente en la comprensión JPEG debido a que es el uso en la gran mayoría de las plataformas digitales por la necesidad de ahorrar costes y tiempo en la transmisión de la foto dentro de la misma. Además, la mayoría de las cámaras dan esta opción por defecto.

El formato JPEG debe su nombre a las siglas G*rupo de Expertos en Fotografía (Joint Photographic Experts Group)*. Cada JPEG es una solo imagen que contiene un solo fotograma que puede contener uno o varios escaneos, dependiendo del modo de codificación [24]. Otro concepto clave sobre los archivos JPEG es que convierten la imagen de entrada de su modelo de color RGB de origen al modelo de color YCbCr, dividiendo la imagen en componentes de brillo, croma azul y croma rojo.

Como se explica brevemente en [20] la compresión JPEG convierte una imagen dada en 8x8 bloques y aplica la transformada de coseno discreta 2D (DCT, *Discrete Cosene Transform*) a cada bloque. Luego, los coeficientes DCT se cuantifican utilizando una tabla de cuantificación

predefinida de 8x8 correspondiente al factor de compresión. El propósito de convertir la imagen es facilitar el filtrado de los cambios de brillo de alta frecuencia que son menos visibles para el ojo humano [24].

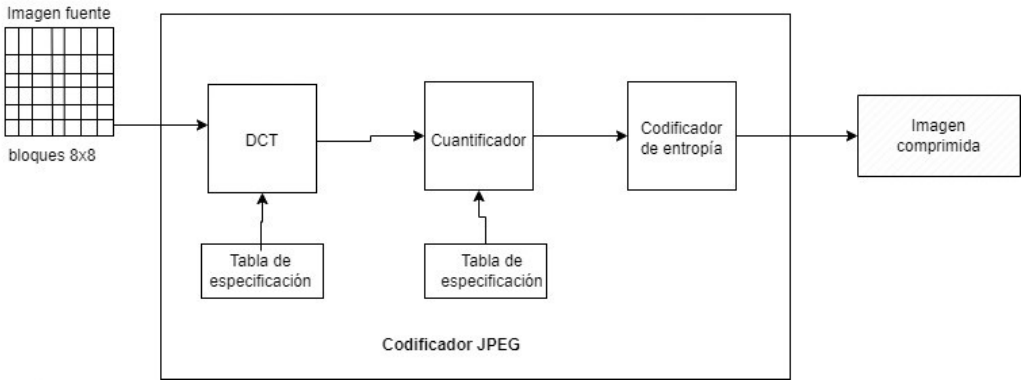

Figura 3: Bloque diagrama codificación secuencial imagen JPEG [50]

La figura 3, sigue el paradigma de codificación de transformación común, según el cual la imagen sufre una transformación lineal invertible, y luego los coeficientes de transformación se cuantifican y codifican por entropía.

SOI	Marcador (1 a n)				SOS			Datos Imagen	EOI
FFD8	FF	No. de Marca (1 byte)	Tamaño de los Datos (2 bytes)	Datos (n bytes)	FFDA	Tamaño de los Datos (2 bytes)	Datos (n bytes)	Datos (n bytes)	FFD9

Tabla 4: Estructura general con marcadores de una imagen JPEG siguiendo el formato EXIF [36]

La estructura de datos A*rchivo de imagen intercambiable* (EXIF, *Exchangeable Image File*), sigue la composición de la tabla 4. Todos los archivos JPEG comienzan con el valor binario de inicio de la imagen 0xFFD8 (SOI, *Start Of Image*) y terminan con valor fin de imagen 0xFFD9 (EOI, *End Of Image*). SOI y EOI son marcadores que no tienen datos posteriores a diferencia de los otros restantes que tienen una estructura fija y datos asociados. El campo con la marca 0xFFDA (SOS, *Strart Of Stream*), indica el inicio de los datos propiamente dichos de la imagen, y precede siempre al campo EOI [36]. Dentro de la zona de marcadores la identificación de cada uno comienza con el campo señalado en la tabla como FF, el cual hace referencia al campo aplicación de marcador 1 (APP1, *Application Marker Segment 1)* e irá desde los valores 0xFFE0~0xFFEF. El

resto de los campos de marcadores corresponde al área de datos de AAP1, incluido el tamaño del descriptor en sí. Algunas técnicas forenses se encargan de extraer los datos EXIF y estudiar la estructura anteriormente mencionada, para dar veracidad a una foto. Este tipo de análisis está ampliamente ligado con el dispositivo que captura la imagen.

3.1.2 Tipos manipulación

Una vez que la imagen está digitalizada, ésta puede someterse a distintas modificaciones para conseguir el efecto deseado, desde añadir elementos que no aparecen en la original, eliminar aquellos que no se quieren visualizar, dar más o menos luz y brillo o reajustar el tamaño, entre otros. Las técnicas más usadas en el retoque fotográfico y manipulación se dividen en:

- Técnica de copiar y pegar
- Cara sintetizada
- Manipulación de atributos
- Técnica de empalme

3.1.2.1 Técnica copiar y pegar

Este sistema consiste en copiar o mover una aparte de una fotografía en otra foto o en parte de la misma [26]. Una de las técnicas quizás más usadas en la falsificación de fotografías y que elaborada con mucho esmero es prácticamente imposible detectar la región tratada. En La figura 4 vemos que en (b) que es la imagen manipulada muestra un cohete demás disparado que no se existe en (a) que es la imagen original.

Figura 4: Ejemplo de técnica de manipulación copiar y pegar en una misma foto [26]

En [25] se realiza un barrido de las manipulaciones faciales que se apoyan sobre esta técnica, las cuales son seguidas por la comunidad de investigadores:

- Intercambio de identidades
- Intercambio de expresiones

3.1.2.1.1 Intercambio de identidades

Esta manipulación consiste en reemplazar la cara de una persona en otra. Este tipo de modificaciones es que se llevó a cabo en la Reddit dando origen a la palabra deepfake. La figura 5 muestra el resultado fake de copiar en la cara original (source) la cara del segundo personaje (target).

Figura 5: Ejemplo manipulación intercambio de identidades [25]

3.1.2.1.2 Intercambio expresiones

Modifica la expresión facial de una persona, para mostrar un gesto o una reacción ante una información o comportamiento. En la figura 6 se puede apreciar un ejemplo de esta modificación donde la cara resultante toma como expresión la cara de la segunda imagen.

Figura 6: Ejemplo manipulación intercambio de expresiones [25]

3.1.2.2 Cara sintetizada

Este tipo de manipulación crea una imagen inexistente de muy alta calidad a partir de una foto de entrada. Esta generación es fruto de potentes tecnologías *Red de Adversario Generativa* (GAN, *Generative Adversarial Network*). La figura 7 muestra un ejemplo donde la imagen etiquetada falsa se genera con características que obtiene a partir de la imagen cargada real, como puede ser el tono de piel, la formas de los ojos o rasgos asociados a la edad.

Figura 7: Cara sintetizada [25]

3.1.2.3 Manipulación en los atributos

Se refiere a ella como la edición o manipulación de algunos atributos propios de la imagen, como el color del pelo, piel, género, edad, etc. Un ejemplo de este tipo de manipulaciones es la aplicación para smartphone FaceApp[4]. La figura 8 muestra una variedad de atributos que se le pueden modificar a la imagen, tales como poner pelo, añadir gafas o corregir imperfecciones.

Figura 8: Ejemplo manipulación tras modificar algunos atributos en la imagen [25]

4 https://www.faceapp.com. Ultimo acceso 16/06/2022

3.1.2.4 Técnica de empalme

Esta modificación se caracteriza por una operación de recorte de una imagen y pegarla en otra imagen, es decir, superponer dos regiones diferentes de dos fotos. El ejemplo que se muestra en la figura 9, se aprecia como un objeto de la foto (a) ha sido recortado y copiado en la foto (b). El resultado se aprecia se aprecia en la figura (c).

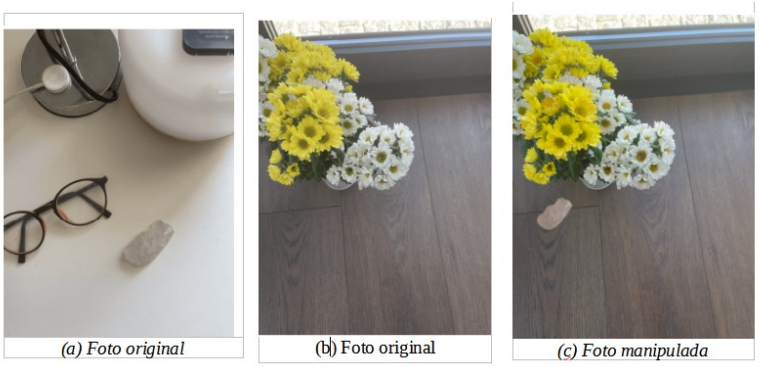

Figura 9: Ejemplo de manipulación con la técnica de empalme [26]

Una vez que el objeto ha sido copiado en otra zona o manipulada, será necesario procesar la imagen destino para suavizar luces, brillos o bordes del objeto, tono de piel, en el caso de manipulaciones faciales, con el objetivo de dar sensación de realidad. Estos procesamientos, hacen que la foto final falsificada tenga atributos diferentes en esas regiones comparado con el resto. Este desfase será aprovechado por los métodos de detección de falsificación.

3.2 Vídeos digitales

Al igual que ocurre con las imágenes, un vídeo es susceptible de sufrir una manipulación. Es importante conocer cómo se forma un vídeo internamente, porque sobre esto versarán las modificaciones por parte del atacante, como saber que tiene que buscar el analista forense, que parámetros son los que se modifican para que la comprensión no sea la que se esperaba o porqué una función se muestra alterada.

Un vídeo se compone de una secuenciada de imágenes o fotogramas con alta calidad [29]. Durante la codificación el grupo de imágenes (GOP, *Group Of Picture)* se especifica el orden de las imágenes. Los distintos tipos de imágenes que puede contener un GOP son:

- Fotograma I (codificación intra) – Es un fotograma de referencia que contiene información sobre el resto de fotogramas a continuación.
- Fotograma P (codificación mediante predicción) – Contiene solo los datos que han cambiado del fotograma anterior (registran el movimiento).
- Fotograma B (codificación mediante predicción bidireccional) -Contienen los datos que han cambiado del fotograma anterior y siguiente, ya sean tipo P o B.

La figura 10 muestra una composición típica de un GOP, la cual siempre empieza por un fotograma tipo I seguida de varios fotogramas P intercalados con fotogramas tipo B. Con el objetivo de reducir espacio y tiempo la señal puede ser comprimida.

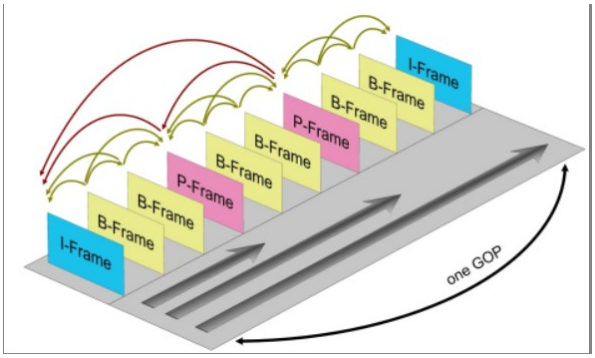

Figura 10: Secuencia de imagen de un vídeo [30]

3.2.1 H.264/MPEG-4 parte 10

H.264 es la compresión de vídeo desarrollada por la Unión Internacional de Telecomunicaciones (como H.264) junto con la Organización Internacional de Normalización / Comisión Electrotécnica Internacional Moving Picture Experts Group (MPEG-4 Parte 10,

Advanced Video Codificación, o AVC) [31]. Este formato de codificación es más eficaz en almacenamiento, rendimiento y transmisión que sus antecesores MPEG-2, H.263 y MPEG-4 parte 2 [32]. Su mayor ventaja es una codificación de alta calidad con menos bits que cualquier otro formato, lo que lo hace ideal para emisiones de televisión, ofreciendo una alta calidad de vídeo con un ancho de banda limitado, y vídeos de alta definición en DVD [31].

La tendencia actual de las empresas, tales como Phillips, Motorola, Broadcom entre otras, es implementar esta codificación aplicándolas a cámaras digitales, teléfonos móviles, transmisión de vídeo sobre redes IP, etc. El estándar se centra únicamente en la codificación del vídeo y son las empresas las que deciden la codificación del audio [32].

3.2.2 Tipos manipulación

Gracias al desarrollo de programas basados en IA al alcance de los usuarios no expertos, es un hecho el auge de la manipulación digital, no solo en imágenes como se ha visto en la sección anterior 3.1.2 , sino en vídeos. A continuación, se realiza un repaso sobre las dos técnicas más generales de manipulación que puede sufrir un vídeo digital. Ambos métodos se basan en la definición de que un vídeo es una secuencia de imágenes, alterar, borrar o insertar los fotogramas serán las bases para su manipulación. En la siguiente figura 11 se visualiza como se van desglosando los distintos tipos de manipulaciones en función del tipo de alteración que sufra el GOP, generando subtipos más específicos.

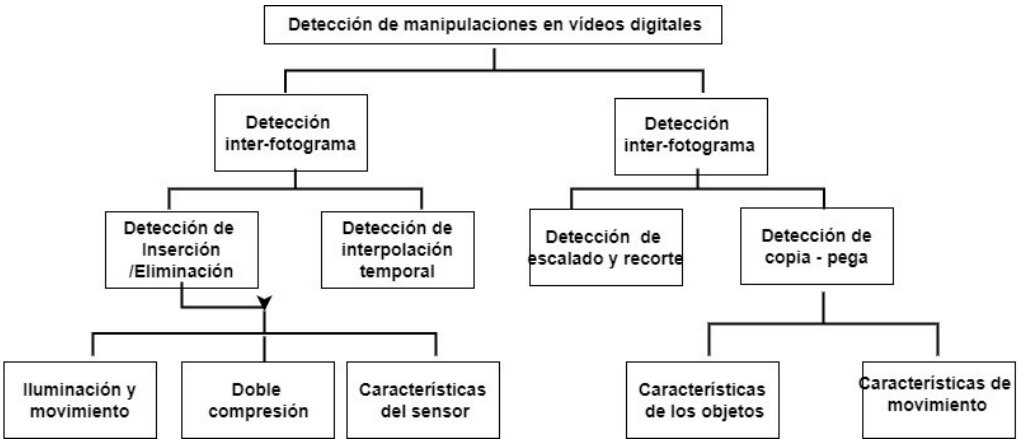

Figura 11: Esquemas de detección de manipulaciones de vídeos [28]

3.2.2.1 Manipulación inter-fotograma

En este tipo de manipulación se centra en la manipulación de la correlación temporal, es posible que los fotogramas han sido eliminados, insertados o duplicados enteramente [28]. La figura 12 muestra las distintas manipulaciones que pueden sufrir un grupo de fotogramas. En la sección (a) estaría representado la imagen original, mientras que en las diferentes secciones vemos estas manipulaciones; En (b) se añaden fotogramas a la secuencia, en (c) se eliminan dos fotogramas centrales y por último en (d) se duplican los 3 últimos fotogramas para añadirlos nuevamente.

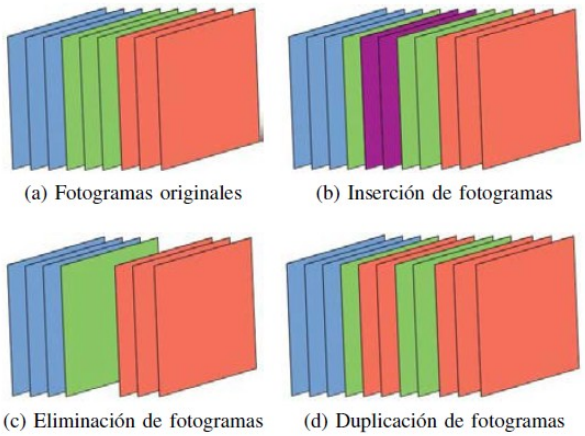

Figura 12: Ejemplo manipulación Inter-fotograma [28]

3.2.2.2 Manipulación intra-fotograma

El atacante altera el contenido de cada fotograma individualmente. En [28] muestra las clasificaciones en la que se puede dividir esta técnica.

- A nivel de pixel. Trata al fotograma como una imagen individual y la manipula con alguna técnica, por ejemplo, las explicadas en la sección anterior 3.1.2 .
- A nivel de fotograma. Cambia el tamaño del fotograma para ocultar cierto contenido que esté en los bordes del fotograma como la hora y lugar de grabación.

3.2.2.3 Doble compresión

Un vídeo sufre una compresión inicial cuando se captura la imagen y una segunda cuando se realiza la falsificación. Este hecho es el fundamento de muchos algoritmos de detección de manipulación de vídeos, puesto que, al detectar más de una compresión, aparece la duda de la falsificación. El análisis forense tiene que encargarse de discernir si la extracompresión es por una falsificación o para ahorrar espacio y tiempo en la carga y descarga de la nube.

3.3 Redes de Adversario Generativas

Cada vez, es más el uso de aplicaciones para la generación de vídeos falsos, debido a que sus interfaces son más amigables y el hecho de que cualquier usuario sin grandes conocimientos puede crear un contenido manipulado, solo con un poco de tiempo, destreza y práctica. Estas aplicaciones hacen uso de las redes de adversario generativas (GAN, *Generarive Adversary Network)*. Aplicaciones tan conocidas como FakeApp[5], OpenFace[6] o Swap Face2Face[7] implementan esta técnica y son capaces de crear caras sintéticas con un nivel muy alto de realismo y detalle [17].

Los algoritmos GAN se componen de dos redes: un **codificador** o **generador** intenta crear una muestra que no se puede distinguir de una real y una red **decodificadora** intenta clasificar las muestras generadas y las reales [37]. Como explica el autor en [37], durante el proceso de entrenamiento se va a producir movimientos continuos entre el generador y el decodificador. Cuando el decodificador es capaz de diferenciar entre las reales y manipuladas, el generador reajusta sus parámetros para producir muestras más reales. A continuación, el discriminador ajustas sus parámetros para diferenciar nuevamente. La figura 13 muestra el esquema de generación de un vídeo falso con GAN, en la que se visualiza una primera parte de codificador, para posteriormente

5 https://randomeo.com/descargas/webapps/descargar-fakeapp/. Último acceso 10/06/2022

6 https://sourceforge.net/projects/openface.mirror/.Último acceso 10/06/2022

7 https://apkpure.com/face-2-face-face-swap/com.atc.gui. .Último acceso 10/06/2022

con las características generadas, se decodifique y dar una aproximación de la imagen. En ella se muestra el proceso de entrenamiento que se repite varias veces.

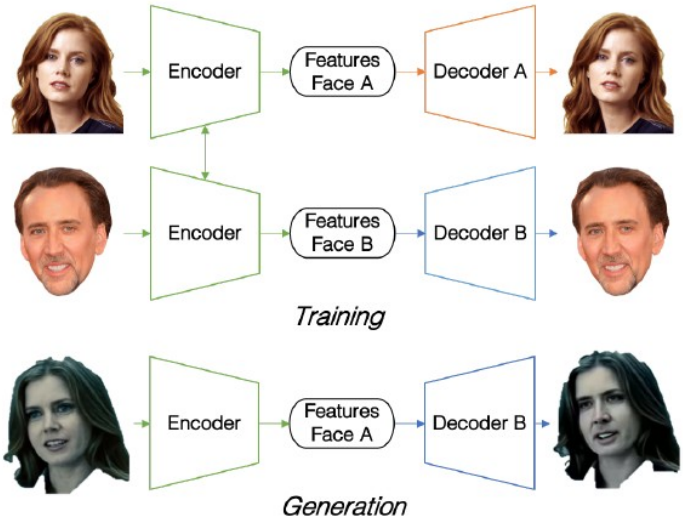

Figura 13: Esquema creación vídeo falso [41]

3.4 Redes Neuronales Artificiales

Las Redes Neuronales Artificiales (ANN, *Artificial Neural Networks*), son sistemas de computación que intentan imitar el diseño en las redes neuronales biológicas. De tal modo que la entrada de una neurona está condicionada por la salida de la precedente. Están formadas por 3 capas interconectadas, una capa de entrada que es la que recibe la información del exterior, una o varias capas ocultas donde se aplican las funciones y algoritmos y la información se va propagando de neurona en neurona y aprendiendo de cada decisión tomada hasta llegar a la última capa de salida, que es la que está interconectada con el mundo exterior y donde se muestra el resultado. En la figura 14 muestra una versión resumida y gráfica del curso de propagación de la información en la que se aprecia que la entrada de cada neurona es la salida de la predecesora hasta dar con la salida final.

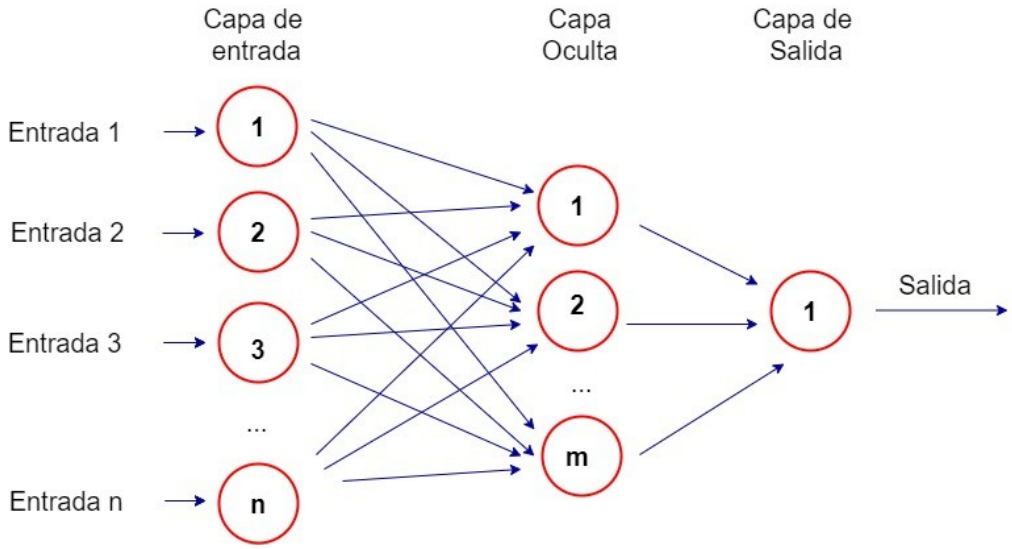

Figura 14: Esquema red neuronal [39]

Cada nodo o neurona tiene una variable de estado, y a cada conexión entre dos nodos se le establece un peso y un umbral asociado. En cada nodo se produce una función de activación de tal modo que si la salida de un nodo es superior al umbral se activa dicho nodo y se envía la salida al siguiente [38]. El hecho de ir pasando información hace que la red sea definida como de propagación hacia delante.

$$\sum_{i=1}^{m} Wi\ Xi + bias \qquad output = f(x) = \left\{ 1\, si \sum W_i X_i + \theta \geq 0 ; 0\, si \sum W_i X_i + \theta < 0 \right\}$$

Cada capa aprende a encontrar y detectar las características que mejor se adapta para clasificar los datos. Este tipo de redes se entrenan para reconozcan patrones, clasifiquen datos y pronostiquen eventos futuros a gran velocidad. El entrenamiento de una red neural se consigue introduciendo datos en la red y la modificación de los pesos según el error conseguido y en función de cuanto haya contribuido cada neurona al resultado de este modo la red aprende y llegando a obtener muy buenos resultados [39].

Existen varios tipos de redes dependiendo de su conexión, el tipo de capas o el grado de conexión.

3.4.1 Redes Neuronales Convolucionales

Este tipo de redes neuronales convolucionales (CNN, *Convolutional Neural Networks*) se usan principalmente para el reconocimiento de patrones de imágenes. Permite codificar diferentes características dentro del red, lo que hace que se reduzcan el número de parámetros para establecer un modelo y por lo tanto más rápidas. De acuerdo al artículo [40], las neuronas en este tipo de redes constan de 3 dimensiones alto, ancho y profundidad.

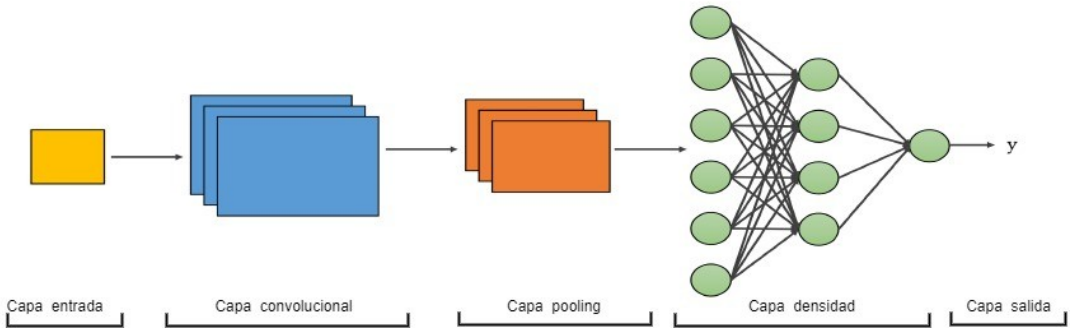

Figura 15: Esquema estructura básica red CNN [51]

Según el esquema de la figura 15, las redes CNN se dividen en 4 áreas:

- Una capa de entrada, que será la entrada de una imagen.
- Una capa convolucional, la cual se encarga de extraer características propias de la imagen y reducirá el tamaño de la misma. De cada entrada generará un mapa 2D de activación a través de las dimensiones espaciales. Estas capas pueden mejorar su complejidad a través de optimizar las salidas. Optimización que se lleva a cabo a través de los parámetros profundidad, el trazo y zero-padding es un proceso de relleno en el borde de la entrada. Estos parámetros alterarán la dimensión espacial de la capa.
- Pulling (conjunto) de capas, ayuda a reducir la dimensión de la representación, el número de parámetros y la complejidad del cálculo. Opera sobre cada mapa de activación y escalando su dimensión usando la función MAX.

- Capa totalmente conectada, las neuronas están directamente conectadas a las dos capas adyacentes sin que ninguna esté conectada entre ellas. Intentan establecer marcas a partir de los mapas de activación que serán utilizadas para la clasificación del resultado.

3.4.2 Redes Neuronales Recurrentes

Las redes neuronales recurrentes (RNN, *Recurrent Neural Networks*) tienen capacidad para secuencia de datos en el tiempo. Utilizada en aplicaciones como Siri de Apple, traductor de Google o búsqueda por voz. Utilizada en el aprendizaje profundo. Al contrario de las CNN, este tipo de redes, la salida de una neurona depende de los elementos previos de la secuencia, lo que lo convierte en una estructura recurrente. La parte izquierda de la figura 16 realiza una representación simple de cómo se compone una RNN y en la parte de la derecha, muestra la misma estructura pero descomprimida y más compleja. En ambos se aprecia que la salida de una neurona necesita más elementos de cálculos para procesar su salida.

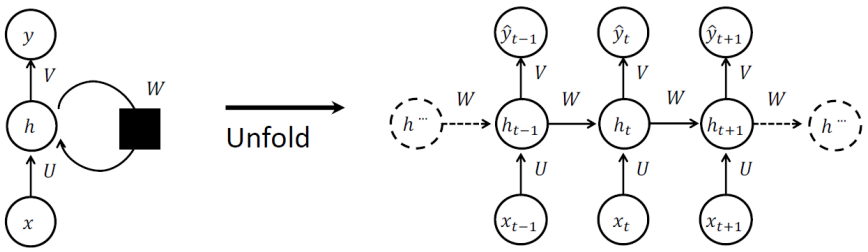

Figura 16: Gráfico RNN desplegado [35]

Como muestra en [35], las RNN consisten en 3 matrices con funciones de activación, tal y como se muestra en la figura 17, que hacen referencia a las fórmulas y elementos que necesitan las neuronas de las RNN vistas en la figura anterior 16.

$$h_t = \tanh(Ux_t + Wh_{t-1})$$

$$\hat{y} = \lambda(Vh_t)$$

- U: input-hidden matrix
- W: hidden-hidden matrix
- V: hidden-output matrix

Figura 17: Representación matrices RNN [35]

Hacen uso del algoritmo de retropropagaciópn (BTT, Back Propagation Trough Time) para determinar los gradientes, el modelo se auto entrena calculando errores desde su capa de salida hasta su capa de entrada [34]. Este proceso presenta dos problemas la explosión y la desaparición de gradiente. Estos problemas se definen por el tamaño del gradiente, que es la pendiente de la función de pérdida a lo largo de la curva de error. Cuando el gradiente es demasiado pequeño, continúa haciéndolo más pequeño, actualizando los parámetros de peso hasta que se vuelven insignificantes.

3.4.3 Long-Short Term Memory (LSTM)

Estas redes con memoria a corto y largo plazo tratan de dar respuesta a los problemas que presentan las RNN, a través de una dependencia a más largo plazo, para ello las células son reemplazadas con células LSTM.

Esta extensión equipa los nodos con celdas de memoria y puertas para controlar el flujo de la información [35], de modo que la información y los gradientes puedan fluir sin cambios entre iteraciones [46].

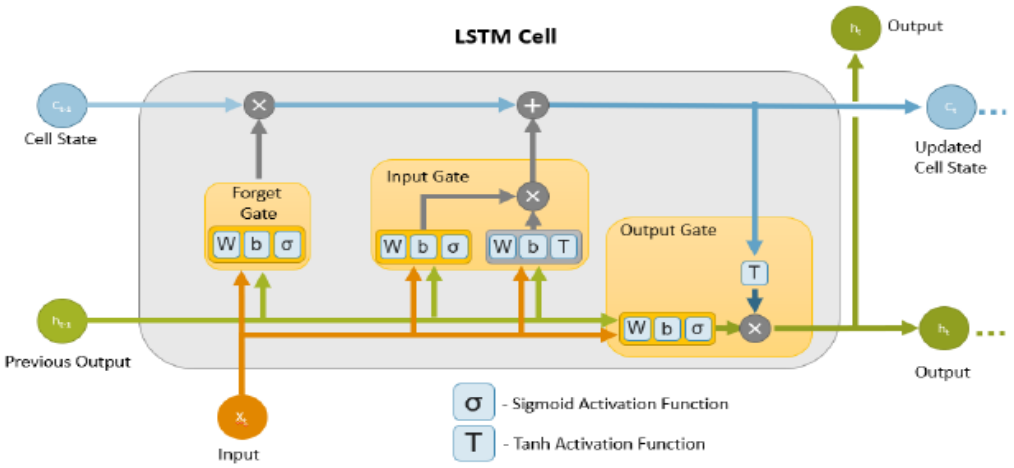

Figura 18: Arquitectura de una celda LSTM [46]

Según se aprecia en la figura 18, la celda consta de 3 puertas que los autores en [46] definen como:

- **Forget Gate (**Puerta de olvido). Decide que información y cuanta se va a eliminar del estado de la celda. La función para activar la puerta es una función sigmoidea; esto asegura que el vector de salida de la puerta produzca valores continuos entre 1 (mantener) y 0 (olvidar).
- **Input Gate.** (Puerta de entrada). Hace uso de dos unidades funcionales. La primera unidad unidad hace uso de la función de activación tanh (tangente hiperbólica) cuyas salidas son valores comprendidos entre -1 y 1 y decide el cambio de estado de la celda. Y la segunda

unidad funcional usa, de nuevo, una función sigmoidea y responde de la magnitud del cambio.

- **Output Gate** (Puerta de salida). Utiliza la matriz entrenada W y el sesgo bo, para obtener información relevante de la entrada actual y la salida anterior, para combinar con el estado actual de la celda para predecir la siguiente salida.

Gracias a la habilidad para el manejo de datos temporales en este tipo de redes es utilizado en aplicaciones que requieren series temporales, como por ejemplo en predicciones de producción de energía solar o funciones de procesamiento de lenguaje natural (NLP, N*atural Language Processing*) tales como asistentes de voz o chatbot, entre otras.

4. Técnicas de análisis forense multimedia

Una distinción fundamental que se hace a menudo en los algoritmos forenses de imágenes es la detección de falsificación de imágenes activa frente a la pasiva. La detección activa de falsificaciones se refiere a algoritmos en los que las imágenes se han procesado de forma preventiva y se han marcado de forma imperceptible, de modo que las alteraciones se puedan detectar y localizar posteriormente. La detección pasiva, por otro lado, trata con imágenes completamente desconocidas sin requisitos previos ni marcas de agua, estudian la imagen a través de las variaciones. De hecho, es esta modalidad la más usada para la detección de falsificaciones en imágenes digitales [27].

4.1 Detección de caras falsas

En [5] los autores presentan una técnica de detección de vídeo falso atendiendo a las características forenses y modificaciones a mano. Con ayuda de la librería dlib[8], antes de procesar la imagen, la cara se segmenta en regiones localizadas en 68 puntos faciales.

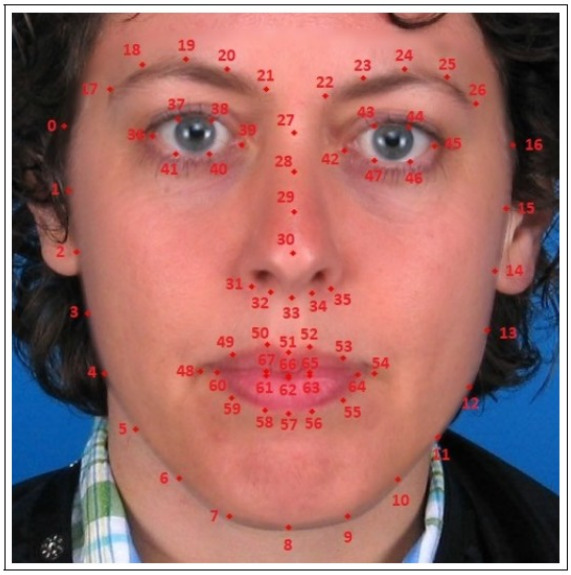

Figura 19: Distribución 68 puntos faciales librería dlib [5]

En la figura 19 se visualiza la distribución de los 68 puntos faciales. El primer conjunto se compone de las marcar del 0 al 26 los cuales describe el borde de la cara desde la frente hasta la barbilla. El segundo conjunto lo compone la región de la nariz desde los puntos 27 al 35. La tercera región son los puntos del 36 al 47 y corresponden a los ojos. Y, por último, la región de la boca que se compone de los puntos del 48 al 67. En el estudio se realiza una combinación de las tres zonas para obtener resultados. Aunque los propios autores demuestran que comparado con una técnica que utilice CNN sus datos son discretamente inferiores, la ventaja que presenta este método es que no requieren un gran dataset (conjunto de datos), por lo que mejora tanto en rendimiento como en recursos, además de la interpretabilidad y las consecuencias que pueda tener para la validación de plausibilidad de decisiones tomadas.

8 http://dlib.net/. Último acceso 10/06/2022

Otro ejemplo en este tipo de detección se muestra en [33] el cual hace uso del modelo DFT-MF, que consiste en extraer fotograma del vídeo, para reducir la gran cantidad de tiempo y energía hacen uso de la herramienta MoviePy[9]. Este modelo se centra en la región de la boca, y al igual que ocurre en el método anterior, hace uso de la librería dlib para realizar las 68 marcas de la región facial como se mostró en la figura 19. En este caso, excluye todos los fotogramas en los que la boca aparece cerrada. El modelo DFTM -MF usará la técnica de aprendizaje profundo (CNN) para clasificar el vídeo como verdadero o falso basándose en un umbral de fotograma falsos identificados en todo el vídeo basado en calcular el número de palabras por frase, velocidad del habla y ratio de fotograma (fps) .

Otra aproximación se basa en el número de palabras por frase en un vídeo. Según el Instituto de Prensa Americano (API, *American Press Institute*) una frase se compone de 8 palabras o menos. La velocidad del habla es la velocidad con a la que habla el orante y se calcula como el número de palabras por minuto (ppm). Los autores señalan en su artículo que existen estudios que muestran que la velocidad de un discurso oficial está entre 120 o 150 ppm, pero un discurso normal depende de varios factores. El número de fps es el número de marcos individuales de una imagen que se muestra por segundo en un vídeo. Los autores establecen un umbral de 5 palabras en una frase. La tabla 5 muestra un pseudocódigo del funcionamiento según la descripción de los parámetros arriba definidos.

Tiempo en segundos	Número de fotogramas	Número de palabras	If (num de fotogragramas falsos en video >50)
			El vídeo es falso;
			else
			El video es real

1	30	2	
2	60	4	
3	90	6	

Tabla 5: Pseudocódigo por número de fotograma falsos [33]

En la evaluación y comparación de este método con otros de la misma naturaleza y mismo conjunto de datos, los autores demuestran que esta técnica obtiene mejores resultados.

9 https://pypi.org/project/moviepy/. Ultimo Acceso 13/06/2022

4.2 Detección de vídeos manipulados a nivel de fotograma

De entre muchas de las técnicas que se pueden utilizar para la detección de manipulaciones tanto en imágenes como en vídeo, los autores en [26] optan por un método basado en máquinas vectoriales de soporte de aprendizaje automático (SVM, *Support Vector Machines*). Dichos autores han desarrollado 3 nuevos módulos implementados en Phyton 3 que han integrado a la herramienta para el análisis forense Autospy[10], dotándola así de más capacidad y funcionalidad. El primer módulo se encarga de extraer 3 de cada 4 fotogramas por segundos de un vídeo de entrada y almacenarlos en el dataset, el segundo método aplica la Transformada Discreta de Founier (DFT, *Discrete Fournier Transform*) a cada imagen y obtener de cada una de ellas 50 características, que serán usadas en el último módulo para estimar si el archivo multimedia es o no falso. La figura 20 muestra los diferentes pasos del proceso, separado en dos procesos; (a) para la fase de pre procesado, donde se extraerían los fotogramas del vídeo y las características DFT de cada imagen con la que se generan los datos de pruebas, y (b) la fase de procesado donde evaluar los resultados obtenidos.

10 https://www.autopsy.com. Último acceso 17/06/2022

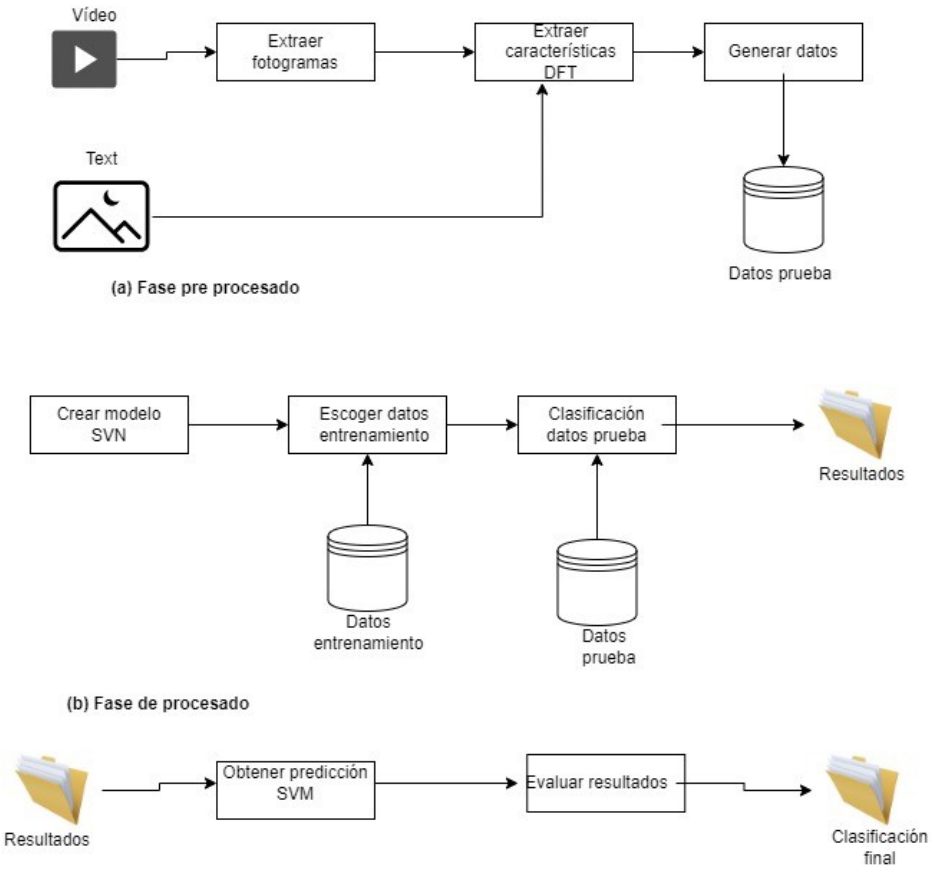

Figura 20: Proceso de los módulos de Autopsy[26]

Se puede observar en el artículo, que los autores con estos módulos implementados no sólo consiguen unos valores bastantes óptimos, tanto con vídeo como con imágenes, sino que realizando la comparación con otros métodos de otros autores basados en CNN, logran unos resultados solo discretamente mejores pero si una significante proyección de tiempo de proceso y de cálculo, además basándose en un dataset más complejo que con los métodos comparados.

En el estudio [41] los autores aprovechan la debilidad que generar las aplicaciones basadas en GAN, visto en el apartado 3.3 , como por ejemplo FakeApp. Esta debilidad se basa en la inconsistencia inter-fotogramas y temporales que se producen a la hora de generar un vídeo falso y que son estudias para determinar la veracidad del mismo. Para la fase del entrenamiento los autores han divido en dos secciones, una primera en la cual usarán un modelo CNN para extraer unos vectores de características por fotograma que servirán como entrada a la segunda parte, que es un

modelo de memoria a corto y largo plazo (LSTM, *Long Short Term Memory)* para analizar la secuencia temporal y discernir si es un vídeo original o falso. La figura 21 muestra un el esquema del algoritmo planteado, como de una entra de fotogramas de una imagen a través del modelo CNN extraen las características para en su segunda fase, evaluar con el modelo LSTM si la imagen no ha sido manipulada o está falsificada.

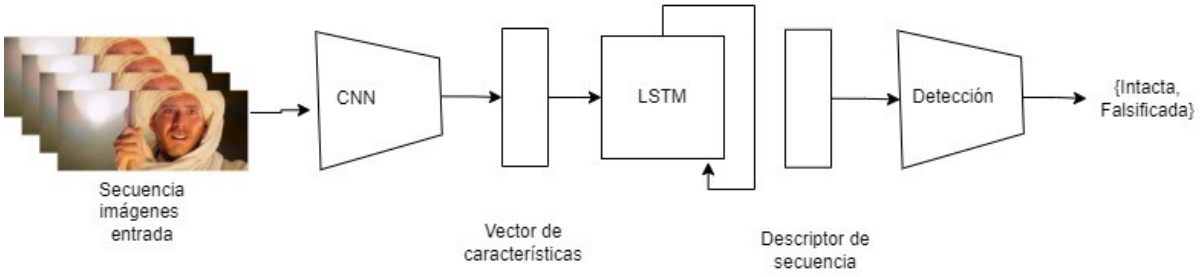

Figura 21: Sistema de detección planteado por los autores [41]

Para la extracción de las características en la primera fase, han adoptado un el algoritmo basado en CCN Inception V3 [42] con las capas completamente conectadas en la parte superior de la red eliminada para generar una representación profunda de cada fotograma utilizando el modelo preentrenado ImagenNet[11]. El vector de características generado de 2048 dimensiones será la secuencia de entrada para la segunda fase, un modelo LSTM. Una vez finalizada la etapa de entrenamiento hacen uso de una capa sofmax para calcular la probabilidad de que el vídeo sea manipulado o no. Según las conclusiones de su estudio en un vídeo de menos dos segundos de vídeo (40 fotogramas para vídeos de muestra a 24 fotogramas por segundo) su método llega a alcanzar una precisión de detección de un 97%.

Aunque el estudio presenta una alta tasa de acierto tanto en los entrenamientos como en la detección faltaría determinar el gasto computacional en tiempo.

11 https://www.image-net.org/ Ultimo acceso 13/06/2022

4.3 Detección de imagen JPEG manipulada

Para el análisis de detección de imágenes JPEG manipuladas, se expone la red de clasificación de manipulación (MCNet, *Manipulation Classification Network*), presentado en [20], el cual trata de explotar las características de domino múltiple del espacio, la frecuencia y la compresión de dominios, basándose en los diferentes tipos de manipulación de las compresiones JPEG. El proceso inicial de entrenamiento los autores lo han dividido en dos fases. La primera fase de entrenamiento propone una técnica de aprendizaje de varias características forenses para múltiples dominios, a través de una estructura múltiple transmisión que explora y aprende de los objetos de visuales y de compresión. Esta estructura se compone de una red de objetos visuales (VANet, *Visual Artifact Network*), con los dominios de espacio y frecuencia causada por las trazas de manipulación, junto con otra red de objetos comprimidos (CANet, *Compression Artifact Network*) para capturar la pérdida de la compresión vía JPEG. Después del entrenamiento, fusionan los mapas de característicos obtenidos de cada una de las redes mencionadas, para aprender las características de dominios múltiples. La siguiente figura 22 muestra una descripción general de la propuesta MCNet. En el prepocesamiento de datos convierte la imagen de entrada en datos de múltiples dominios. La disposición de los bloques de la imagen ha sido alterada con respecto a la original para dar mayor visualización a los textos.

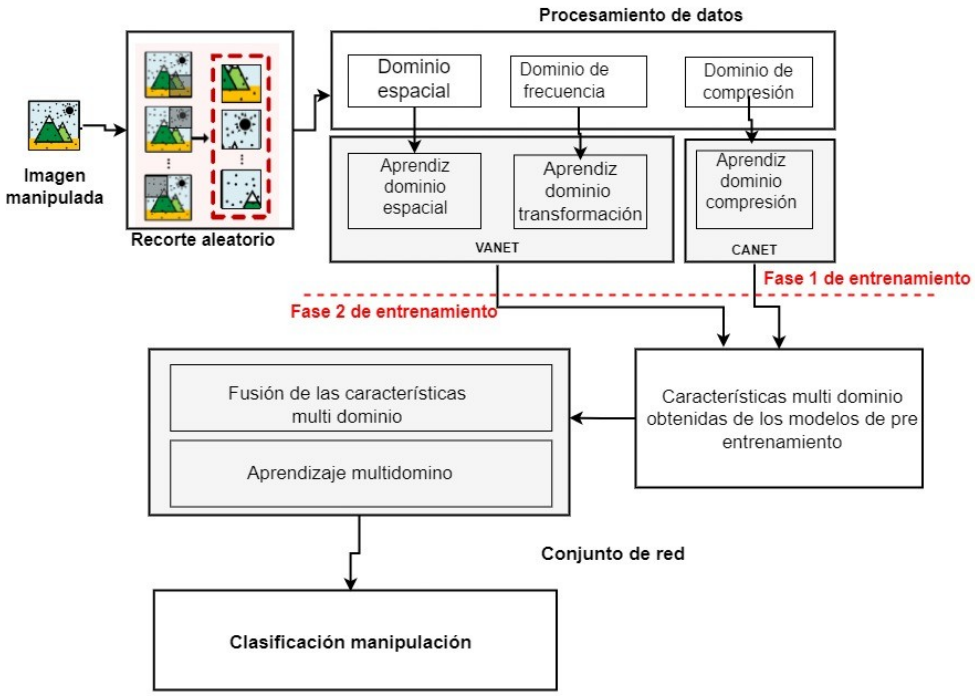

Figura 22: Esquema algoritmo MCNet [20]

La red VANet hace uso de datos procesados basados en los pixeles RGB decodificados a partir de los coeficientes DCT cuantificados. CANet usa la información estática tal como el histograma DTC es usada para detectar doble compresión JPEG. Para comparar sus resultados con otros algoritmos, los autores se fijan en la tasa de error de los mismos. Sus resultados muestran la tasa de error más baja de los algoritmos comparados. La tabla 6 muestra un resumen con los resultados de los modelos de redes expuestas, mientras que los mismos resultados se pueden observar de manera gráfica en la figura 23 donde visualmente se puede apreciar que el modelo MCNET es el que contempla la tasa de errores menor.

Modelo	Clasificación de Errores (%)		
	Error Top1	Error Top2	Diff (%)
VANEt	15.97	4.45	11.52
CANet	36.29	16.17	20.12
MCNet	**15.21**	**4.13**	**11.08**

Tabla 6: Tabla de resultados de comparación [20]

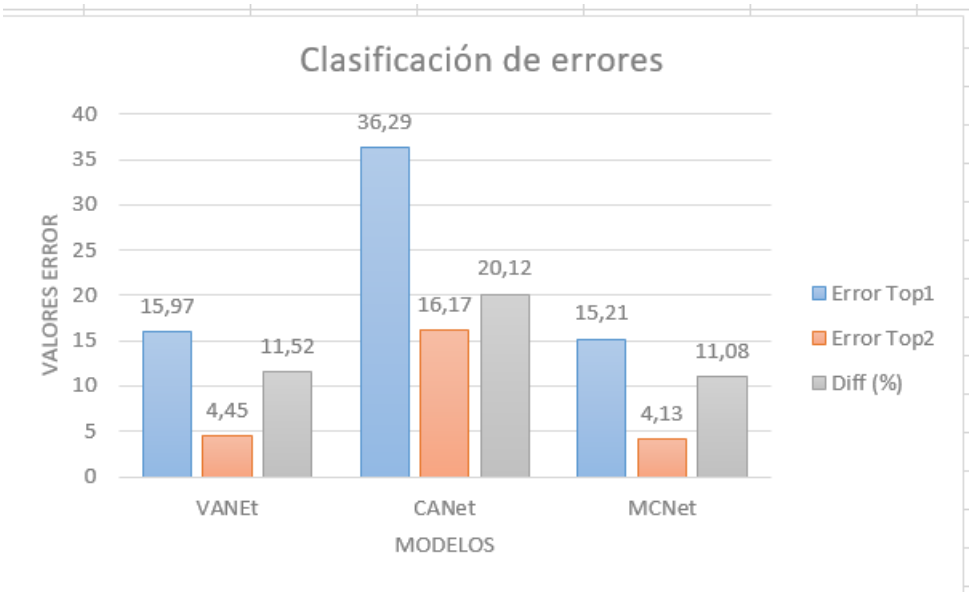

Figura 23: Comparación gráfica errores tabla 6

5. Experimento propio del trabajo

Los trabajos que a continuación se exponen, han sido ejecutados en sobre un Ubuntu 20 de 64 bits virtualizado con 2 CPU sobre Windows 10, AMD Ryzen 7 3700U con Radeon Vega Mobile Gfx 2.30 GHz y 16 Gb de RAM. Las arquitecturas propuestas han sido implementadas con Python 3.8.10, keras 2.8.0 y Numpy 1.22.2.

Esta sección se expone los resultados de los experimentos de las dos arquitecturas *Meso-4 y MesoInception-4* para el estudio de falsificaciones digitales. Para la realización de los mismos se sigue la metodología desarrolla en [43], donde los autores se centran en la detección de caras falsificadas en vídeos, adoptando una aproximación intermedia de una red CNN, vista anteriormente en el apartado 3.4.1 , con un número de capas reducidas. Sobre estos modelos se estudian 4 tipos de conjunto de imágenes:

- Verdadero positivo (VP). Son aquellas imágenes que el algoritmo ha etiquetado correctamente como reales sin ninguna manipulación.
- Verdadero negativo (VN). Son aquellas imágenes que el algoritmo ha evaluado correctamente como manipuladas.
- Falso positivo (FP). Son aquellas imágenes que el algoritmo ha evaluado erróneamente como como reales pero que han sido modificadas.

- Falso negativo (FN). Son aquellas imágenes que el algoritmo evalúa erróneamente como falsas, pero no han sufrido modificación por lo tanto son reales.

Los autores consiguieron los mejores pesos de optimización de la red, en ambos modelos, con lotes de 75 imágenes de 256 x 256 x 3, correspondientes a las dimensiones alto x ancho x número de canales para los colores (RGB) usando ADAM [52] con los parámetros por defecto (β_1=0.9 y β_2=0.999).

En la figura 24 se refleja el proceso que sigue la imagen para el desarrollo. En la primera parte refleja el proceso de extraer los fotogramas del vídeo y procesar las imágenes para que después, Mesonet en cada capa convolucional fije el tamaño y el número de filtros. Filtros que representan una característica distinta de una imagen. Durante la convolución el filtro se pasa sobre toda la imagen que identifica la existencia y la localización de una característica específica. Después en la capa de normalización, cada entrada en la capa reduce la interdependencia entre los parámetros de una capa y la entrada de distribución de la siguiente. En la última capa, es donde la dimensión del dato queda reducida de manera significante, dando así velocidad en los cálculos. MesonNet usa la función max para obtener los valores máximos de cada región de pixel. En resumen, en cada capa convolucional, el modelo es capaz de localizar las características más importantes.

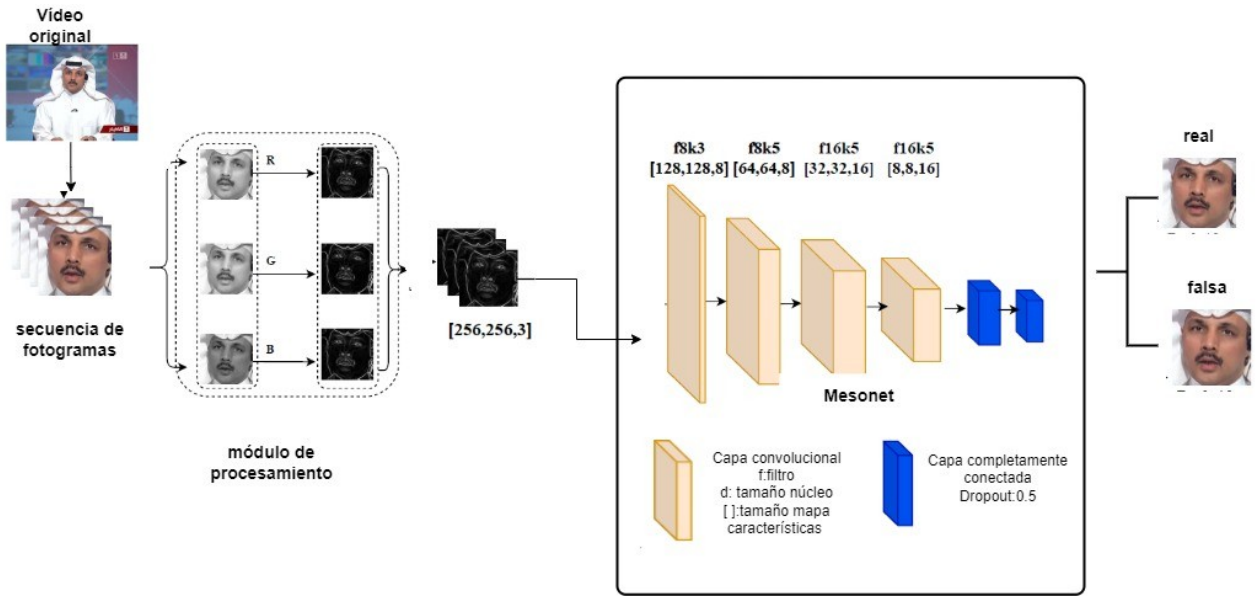

Figura 24: Resumen gráfico metodología MesoNet [53]

5.1 Conjunto de datos

Para el estudio de las arquitecturas, los autores se han creado sus propios dataset convenientemente separados para el entrenamiento y los test. Para ello, han usado dos técnicas conocidas como Deepfake, la cual remplaza la cara de una persona por otra en un vídeo, técnica muy mejorada en la aplicación FakeApp, y el método de Face2Face (F2F), que transfiere la expresión de una cara en otra persona. Las figuras 25 usando la técnica Deepfake se aprecia el cambio de cara, provocando un efecto similar al explicado en el apartado 3.1.2.3 en el que se modifican los atributos, en este caso la forma de los ojos y las expresiones de la cara y en 26 esta vez con la técnica de Face2Face, han intercambiado la expresión de la cara de la imagen original, tal y como se indica en el apartado 3.1.2.1.2 .

Figura 25: Ejemplo de imagen usando técnica Deepfake [43]

Figura 26: Ejemplo de imagen usando técnica Face2Face [43]

Para el conjunto de imágenes con Deepfake los autores recopilaron hasta 175 vídeos falsificados y el doble el conjunto de datos reales de diferentes plataformas, con un rango de duración de 2 a 3 minutos y con una resolución de 854x480 pixeles y con diferentes niveles de compresión usando la codificación H.264. Para la obtención de las caras se han basado en el detector Viola-Jones [47]. El número de fotogramas obtenidos es proporcional al número de ángulos de la cámara, aproximadamente obtuvieron 50 fotogramas por escena.

Para formar el conjunto de datos de Face2Face, los autores recopilaron 300 vídeos del conjunto de datos FaceForensic [48] la cual contiene una extensa recopilación de vídeos usando esta técnica y la cual se encuentra ya divida correctamente en entrenamiento y test. Los vídeos fueron comprimidos igualmente en H.264. El resultado del conjunto total de imágenes se muestra en la tabla 7.

Conjunto Datos	Falsas	Reales
Deepfake entrenamiento	5111	7250
Deepfake pruebas	2889	4259
Face2Face entrenamiento	4500	4500
Face2Face pruebas	3000	3000

Tabla 7: Conjunto de datos totales usados por los autores

El conjunto de datos con el que se lleva a cabo el experimento, es una copia facilitada por los autores y disponible públicamente[12]. Para replicar el trabajo ha sido necesario reducir el número de imágenes de la muestra por motivos técnicos del equipo utilizado. De esta forma se optimiza la capacidad del equipo en el desarrollo de los algoritmos y se facilita la gran cantidad de cálculos necesarios para el tratamiento de tal volumen de imágenes. En un ejercicio de prueba y error, se ha ido reduciendo equitativamente el conjunto de datos hasta encontrar el tamaño óptimo considerando los tiempos de cálculo. Finalmente el conjunto de datos corresponde al 50 % de la muestra de la parte de pruebas original. Los resultados se muestran en la tabla 8 clasificadas en las categorías de falsas y reales.

	Falsas	Reales	Total
Entrenamiento	3603	3700	7303
Validación	2845	4259	7104

Tabla 8: Conjunto de datos propio

12 /h/tps://github.com/DariusAf/MesoNet. Último acceso 14/06/2022

5.2 Meso-4

La red comienza con una secuencia de 4 capas convolucionales seguidas de un proceso de normalización y agrupaciones, seguido de una capa oculta completamente conectada Para mejorar la generalización, las capas convolucionales usan la activación ReLU[13] para regular su producción y evitar la desaparición efecto degradado, y en las capas totalmente conectadas utilizan Dorpout [44] para regularizar y mejorar su robustez. La figura 27 muestra más detallada la parte Mesonet de la figura 24.

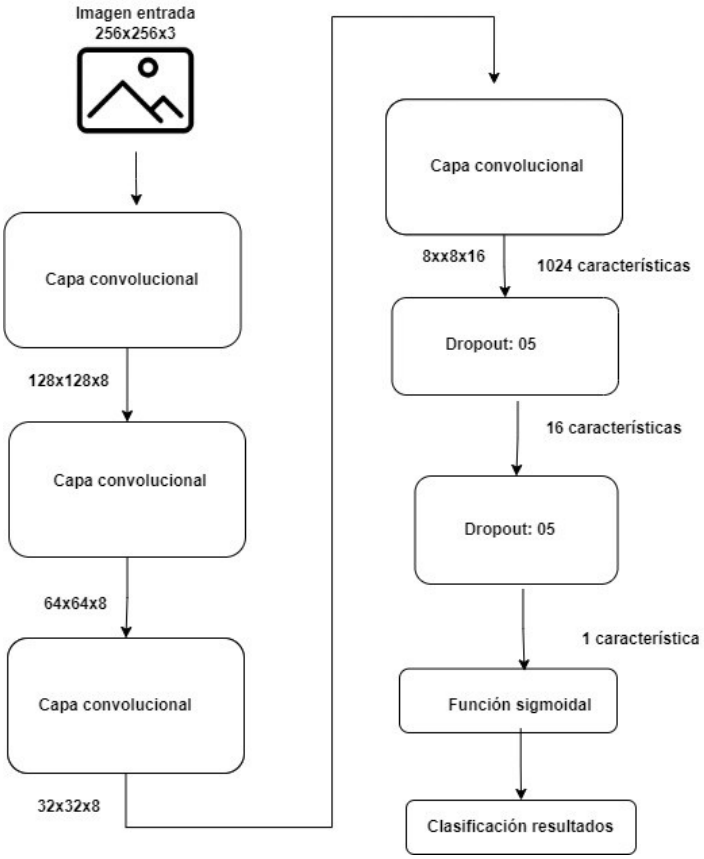

Figura 27: Arquitectura Meso-4 [43]

13 https://deepai.org/machine-learning-glossary-and-terms/relu. Último acceso 14/06/2022

5.3 MesoInception-4

Esta arquitectura es una variante de la anterior, en la que los autores han reemplazado las dos primeras capas convolucionales del modelo anterior, mediante un *módulo de inception.* Unifican la salida de varias capas con diferentes núcleos para así aumentar el espacio de funciones con el que se optimiza el modelo. La figura 28 muestra los detalles de la estructura. La imagen de entrada sigue el mismo patrón de funcionamiento que la figura 24.

.

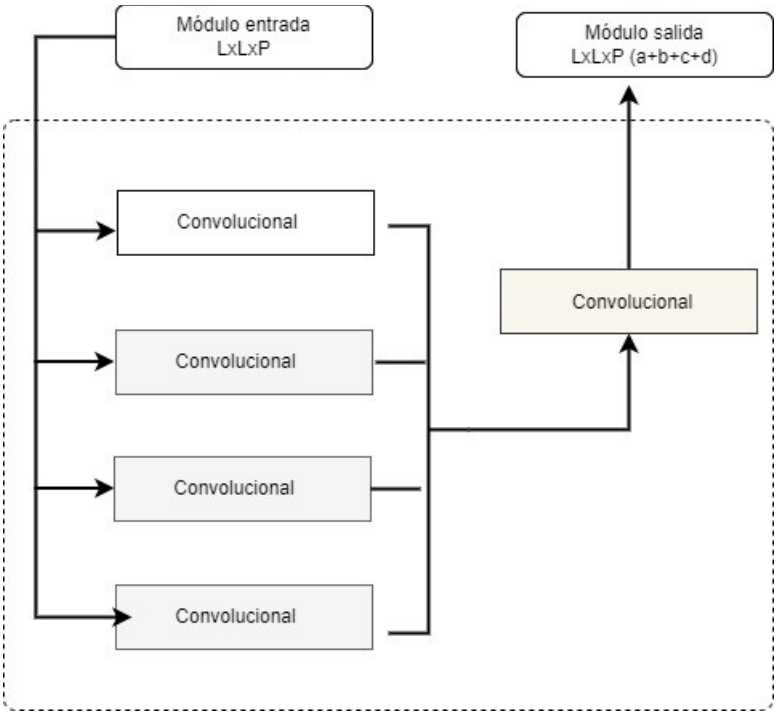

Figura 28: Arquitectura MesoInception-4 [43]

5.4 Clasificación de los resultados

Ambos algoritmos ponderan cada una de las imágenes del conjunto de datos, de manera que si el valor es próximo a 1 consideran la imagen como real, por el contrario, cuanto más cercano se encuentra el valor en 0 la imagen es considerada como falsa. Basada en esta premisa y con el conjunto de datos final después de la reducción, la tabla 9 muestra un resumen de los resultados obtenidos, en ella se muestran todos los valores de las distintas clasificaciones de las imágenes (VP, VN, FP y FN).

Modelos	**Deepfake**				**F2F**			
Arquitectura	**VP**	**VN**	**FP**	**FN**	**VP**	**VN**	**FP**	**FN**
Meso 4	3746	2563	514	281	3884	639	375	2206
Meso Inception 4	2832	2654	427	191	3892	385	368	2459

Tabla 9: Resultados del experimento

Para corroborar el rendimiento de los métodos propuestos, a continuación, se detallan una serie de métricas que lo evidencian, para el cálculo de ellas hay que fijarse en los datos obtenidos que muestra la tabla 9.

- **Precisión (P).** Es la relación entre las predicciones positivas correctas con respecto a todos los casos que se han predicho positivamente Su fórmula queda definida de la siguiente manera:

$$P = {VP} \Big/ {(VP + FP)}$$

Según la fórmula establecida, los resultados para la precisión se muestran en la tabla 10.

Modelo	Deepfake	F2F
Meso 4	0,88	0,91
MesoInception 4	0,90	0,91

Tabla 10: Valores de precisión

Se comprueba que el comportamiento de los algoritmos es el similar, puesto que arroja unos valores muy parecidos independientemente del modelo a usar. Hay una mínima variación inapreciable con Meso 4 en el modelo Deepfake con respecto a MesoInception 4. Los valores obtenidos muestran unos resultados de precisión bastante altos y buenos.

- **Sensibilidad (R, *Recall*).** Es la relación entre las predicciones positivas correctas con respecto a todos los casos que en realidad si son positivos. Su fórmula queda definida de la siguiente manera:

$$R = {VP} \Big/ {(VP + FN)}$$

Según la fórmula establecida, los resultados para la sensibilidad se muestran en la tabla 11.

Modelo	Deepfake	F2F
Meso 4	0,93	0,64
MesoInception 4	0,95	0,61

Tabla 11: Valores de sensibilidad

Al igual que ocurría con la precisión, los algoritmos nuevamente vuelven a mostrar unos resultados similares aunque MesoInception 4 supera en dos décimas a Meso 4. No obstante, si que se aprecia que ambos dan mejores resultados con Deepfake, en la que se obtienen unas tasas muy altas pero que son algo discretas con el modelo F2F.

- **Exactitud.** Es la relación de todas las predicciones correctas con respecto al número total de predicciones. Su fórmula queda definida de la siguiente manera:

$$Exactitud = {(VP+VN)} \Big/ {(VP+VN+FP+FN)}$$

Según la fórmula establecida, los resultados para la exactitud se reflejan en la tabla 12.

Modelo	Deepfake	F2F
Meso 4	0,89	0,64
MesoInception 4	0,91	0,60

Tabla 12: Valores de exactitud

El patrón de comportamiento se repite entre los algoritmos, obteniendo unos valores muy parecidos y con pequeñas diferencias. Se observa que, aunque Meso 4 con el modelo Deepfake es 2 décimas inferior con respecto a MesoInception 4; Con el cambio de modelo ocurre lo contrario, en este caso Meso 4 muestra 4 décimas mejor que MesoInception 4.

- **F1 Score.** Establece la medida armónica entre la precisión y sensibilidad. Su fórmula queda definida definida de la siguiente manera:

$$F1 = 2 \times \left({P \times R} \Big/ {P+R} \right)$$

Para calcular los datos según la fórmula establecida, hay que fijarse en los valores obtenidos en las métricas de precisión (tabla 10) y de sensibilidad (tabla 11). Los resultados para el F1 score se muestran en la tabla 13.

Modelo	Deepfake	F2F
Meso 4	0,90	0,75
Meso Inception 4	0,93	0,73

Tabla 13: Valores de F1 score

Se constata nuevamente el comportamiento que se ha venido repitiendo en las anteriores métricas, en el que se aprecian unas variaciones de décimas entre modelos. En esta métrica, se consiguen unos valores altos para el modelo Deepfake 0,90 para Meso 4 y MesoInception 0,93, por el contrario, con el modelo F2F los valores se vuelven más discretos aunque siguen siendo valores aceptables.

Con todas las mediciones realizadas, y con el conjunto de datos propuesto, las conclusiones que se pueden obtener es que no existe gran diferencia en los resultados entre el uso de un algoritmo u otro sobre los diferentes modelos. El algoritmo que mejora con un modelo luego invierte el resultado con el cambio del modelo. Se destaca que MesoInception 4 con el modelo Deepfake consigue discretamente unos datos mejores que Meso 4. El algoritmo Meso 4 muestra mejores resultados que MesoInception4 usando el modelo F2F. Con estos resultados, no existe una brecha suficiente para discernir cuál de los dos algoritmos funciona mejor. No obstante, en términos generales y fijándose en resultados obtenidos en F1 score (tabla 13), se podrían valorar los algoritmos como un buen mecanismo de detección de vídeos falsificados.

Modelo	Deepfake	F2F
Arquitectura	Reales (VP)	Reales (VP)
Meso 4	0,901	0,946
Meso Inception 4	0,934	0,968

Tabla 14: Registros obtenidos por los autores [43]

La tabla 14 muestra los valores que los autores han obtenido tras realizar sus validaciones. La comprobación de sus resultados con los del experimento propio se hace complicada puesto el

conjunto de imágenes con los que se ha trabajado es bastante más reducido, No obstante si comparamos sus valores con la tabla F1 score (13), en ambos algoritmos con el modelo Deepfake si que se muestran los mismos porcentajes, cosa que no ocurre con ninguno de los algoritmos con el modelo F2F, puesto que los valores de los autores son bastante elevados, incluso rozan un 97% y los del experimento se quedan en marcas más discretas en el que la marca máxima ha sido del 75% con Meso4.

6. Conclusiones y Trabajos futuros

6.1 Conclusiones

En la introducción de este trabajo fin de máster, se ha realizado una exposición sobre cómo afectan las deepfakes en nuestro entorno, hasta el punto de llegar a influir en unas elecciones generales como se apreció en las pasadas elecciones del 2016 en Estados Unidos, y el daño que pueden causar hacia una persona, hecho que se pone de manifiesto en los estudios [45] y [10] donde indica que la mayoría de estas manipulaciones son sobre mujeres y de contenido pornográfico. La proliferación de este fenómeno es debido, entre otras cosas, al abaratamiento de la tecnología, continuamente se van desarrollando nuevos modelos de teléfonos inteligentes provistos de más capacidad de almacenaje y procesamiento. Es importante, además, tener presente que los programas de retoque o manipulación igualmente son más asequibles e incluso gratis en sus versiones de prueba, que aunque internamente están desarrollados con unas metodologías de computación complicadas y potentes como se han expuesto en el apartado teórico, en el manejo de su interfaces son sencillas e intuitivas para usuario, lo que hace que no se necesiten grandes conocimientos, únicamente dedicarle tiempo. Así mismo, se ha indicado que el uso de nuestros dispositivos móviles es el acceso de información de muchos usuarios [3], y en su mayoría a través de redes sociales, en la que existe una evidente falta de regulación legislativa en cuanto al uso y contenido y que

contribuye a la generación y distribución de la deepfake, puesto que el usuario confía en el emisor, bien porque es un conocido o porque inconscientemente damos por válida una noticia de la que ideológicamente estamos de acuerdo y no desconfiamos [8]. Sin embargo, estos avances no solo contribuyen de manera negativa en nuestras vidas, la industria cinematográfica y televisiva también se ha visto beneficiada y aporta nuevas formas creativas de trabajo, como el citado anuncio de una conocida marca de cerveza [14].

Del mismo modo que una persona que genera contenido malicioso se sirve del acceso gratuito de imágenes a través de entrevistas, programas, vídeos de YouTube, perfiles abiertos en redes sociales, etc. donde se suben gran cantidad material audiovisual de gran calidad para generar la deepfake, proporciona un conjunto de datos óptimo para poner a prueba los algoritmos que ayuden a detectar falsificaciones, y ésta, es la premisa que han utilizado los autores en la exposición de su metodología MesoNet, que es la que se ha replicado en el experimento propio, donde de cada vídeo de una duración de 2 a 3 minutos han obtenido un número de fotogramas combinando los ángulos y la iluminación. La arquitectura propuesta realiza una adaptación usando una CNN y basada en el análisis de vectores de características hace que sea un sistema ideal para la detección de caras falsificadas.

Como se ha mencionado al principio del trabajo, el trabajo de un analista forense para la detección de imágenes falsificadas consta de la identificación de la fuente con la que se realiza la captura de la imagen, la cual queda fuera del ámbito de este proyecto, y una fase de autentificación. En esta segunda tarea, es necesario que el analista forense conozca la composición del formato y la compresión de un vídeo, sus características básicas, para que durante el análisis de una evidencia sea capaz de reconocer la ausencia o alteración de una característica que darán lugar a las manipulaciones más extendidas, así como las distintas metodologías, algunas que se incluyen en una herramienta de sobrado conocimiento para el analista como Autopsy[10] y que los autores avalan muy buenos resultados. No obstante, el método propuesto en el experimento propio, además de conseguir muy buenos resultados sin un coste de ejecución elevado, cuenta con el aditivo de que no se requiere un gran equipo para su ejecución, lo que hace que sea una herramienta más a tener en cuenta para el análisis forense.

6.2 Trabajos futuros

Como se ha comentado en el apartado 5.1 el juego de imágenes para la elaboración del proyecto hubo de ser reducido debido a la falta de capacidad del equipo para procesar los cálculos con un número tan elevado de imágenes, hecho por el que no se pudo hacer una demostración 100% del trabajo de los autores. Aunque los valores obtenidos en el experimento propio son altos y no distan mucho de los originales, los autores aseguran, y en sus datos se confirma, que su efectividad es mayor. Por lo tanto, se propone como trabajo futuro trasladar el entorno de desarrollo a una tecnología más potente y menos limitada como en la nube y trabajar con Big Data para poder llevar a cabo la demostración con el conjunto completo facilitado.

A lo largo de este trabajo han sido expuestas una amplia gama de manipulaciones, tanto de vídeo como de imágenes, sin embargo, ha quedado fuera de este estudio la identificación de la fuente con la que se captura la imagen, proceso de gran peso sobre todo en de cara a autentificar evidencias en un procedimiento legal. Esta identificación se puede realizar gracias a los metadatos que se guardan junto con toda la información de la imagen. Cada dispositivo tiene unas características que dejan huella, de tal modo que la ausencia de estas o la alteración, podría considerarse como una manipulación de la misma. La extracción y el estudio de los datos EXIF, sería una línea abierta de investigación junto con las técnicas y propuestas para evitar que estas características fueran manipuladas.

También, se ha puesto de manifiesto la veracidad que los usuarios dan al contenido visual que reciben, por la confianza que depositan en el emisor, por diferentes motivos, políticos, religiosos, etc. Muchos de los vídeos manipulados, la única alteración que han sufrido ha sido una perturbación en el audio para cambiar el discurso y así causar rechazo o polémica sobre la persona que lo está relatando. El fenómeno es conocido como discursos sintéticos que logran que el mensaje parezca real. Para ello, utilizan técnicas basados en aprendizaje profundo y alimentado con horas de conferencias y charlas de esa persona con el objetivo de que el algoritmo aprenda a imitar al interlocutor y dar mayor credibilidad a la falsificación. El ámbito de estudio sobre este campo se propone para una segunda fase en la detección de deepfakes.

Bibliografía

[1] Observatorio Nacional de Tecnología y Sociedad. 2021. Usos y actitudes de consumo de contenidos digitales en España. Madrid: Ministerio de Asuntos Económicos y Transformación Digital. https://www.mineco.gob.es/stfls/mineco/ministerio/ficheros/libreria/UsosDigitales_PDF.pdf. Último acceso 17/06/2022.

[2] Podcats Código de barras. 2021, Noviembre. Twitter, el periódico de los adolescentes. https://cadenaser.com/programa/2021/11/28/codigo_de_barras/1638086453_798572.html. Último acceso 16/06/2022.

[3] Patil, U., & Chouragade, P. M. 2021, August. Deepfake Video Authentication Based on Blockchain. In 2021 Second International Conference on Electronics and Sustainable

[4] Jan Kietzmann, Linda W. Lee, Ian P. McCarthy Tim C. Kietzmann. Deepfakes: Trick or treat?.Business Horizons 2020. 63, 135e146 Communication Systems (ICESC) (pp. 1110-1113). IEEE.

[5] Siegel, D., Kraetzer, C., Seidlitz, S., & Dittmann, J. 2021. Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features. Journal of Imaging, 7(7), 1.

[6] Varela, J. (2005). El asalto de los medios sociales. Cuadernos de periodistas,2, 20-34.

[7] García-Ull, F. J. 2021. Deepfakes: el próximo reto en la detección de noticias falsas. Anàlisi 64, 103-120.

[8] Souza Freire P M, Matias da Silva F r , Ribeiro Goldschmidt g. Fake news detection based on explicit and implicit signals of a hybrid crowd: An approach inspired in meta-learning, Expert Systems with Applications,Volume 183,2021,115414, ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2021.115414. https://www.sciencedirect.com/science/article/pii/S0957417421008344. Último acceso 17/06/2022.

[9] Marwick, A., & Lewis, R. 2017. Media manipulation and disinformation online. New York: Data & Society Research Institute, 7-19.

[10] Cerdán Martínez, V.; Padilla Castillo, G. 2019. Historia del fake audiovisual: Deepfake y la mujer en un imaginario falsificado y perverso, en Historia y comunicación social 24 (2), 505-520.

[11] Allcott H, Gentzkow M.Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 211–236.

[12] The 2013 World Economic Forum's Global Risk Report highlighted the risk of digital misinformation and connected it directly to cyberattacks Howell, L. (Ed.). 2013. Global risks 2013. World Economic Forum.

[13] Westerlund, M. 2019. The emergence of deepfake technology: A review. Technology Innovation Management Review, 9, 40-53.

[14] Palomo-Domínguez, I. 2021. Del mito a la viralidad. El caso de la campaña de Cruzcampo que resucitó a Lola Flores. aDResearch ESIC 26, e262 https://doi.org/10.7263/adresic-026-02. Último acesoo 17/06/2022.

[15] Zampoglou, M., Papadopoulos, S., & Kompatsiaris, Y. 2017. Large-scale evaluation of splicing localization algorithms for web images. Multimedia Tools and Applications,76(4), 4801-4834.

[16] Koopman, Marissa & Macarulla Rodriguez, Andrea & Geradts, Zeno. 2018. Detection of Deepfake Video Manipulation.

[17] L. Verdoliva, Media Forensics and DeepFakes: An Overview. in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910-932, Aug. 2020, doi: 10.1109/JSTSP.2020.3002101.

[18] Quinto, Carlos & Alejandro, Esteban & Armas Vega, Esteban & Lucila, Ana & Sandoval Orozco, Ana & Javier, Luis & Villalba, García & Análisis, Grupo. 2016. Análisis de metadatos en vídeos digitales de dispositivos móviles. VIII Congreso Internacional de Computación y Telecomunicaciones 2016.

[19] Lanh, Tran & Chong, Kai-Sen & Emmanuel, Sabu & Kankanhalli, Mohan. 2007. A Survey on Digital Camera Image Forensic Methods. Proceedings of the 2007 IEEE International Conference on Multimedia and Expo, ICME 2007. 16 - 19. 10.1109/ICME.2007.4284575.

[20] Yu, I. J., Nam, S. H., Ahn, W., Kwon, M. J., & Lee, H. K. (2020). Manipulation classification for jpeg images using multi-domain features. IEEE Access, 8, 210837-210854

[21] Bayar, B., & Stamm, M. C. 2017, September. Augmented convolutional feature maps for robust cnn-based camera model identification. In 2017 IEEE International Conference on Image Processing (ICIP) (pp. 4098-4102). IEEE.

[22] Kok, Chi-Wah & Tam, Wing-shan. 2019. Digital Image Interpolation in Matlab. 10.1002/9781119119623. https://learning.oreilly.com/library/view/digital-image-interpolation/9781119119616/c02.xhtml. Último acceso 08/06/2022.

[23] Introducción a procesamiento digital de imágenes con Matlab. https://www.unioviedo.es/compnum/laboratorios_web/laborat03_intro_imagen/laborat03.html. Último acceso 17/06/2022.

[24] Barni, M. 2018. Document and Image compression. CRC press. https://learning.oreilly.com/library/view/high-performance-images/9781491925799/ch04.html#idm139840163299680. Último acceso 08/06/2022.

[25] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion,64, 131-148.

[26] Ferreira, S., Antunes, M., & Correia, M. E. 2021. Exposing Manipulated Photos and Videos in Digital Forensics Analysis. Journal of Imaging,7(7), 102.

[27] Mishra, M., & Adhikary, F. 2013. Digital image tamper detection techniques-a comprehensive study. arXiv preprint arXiv:1306.6737.

[28] Fernández, E. G., Orozco, A. L. S., & Villalba, L. J. G. Técnica de Detección de Manipulación en Vídeos Digitales Basada en los Algoritmos de Compresión.

[29] Gironi, A., Fontani, M., Bianchi, T., Piva, A., & Barni, M. 2014, May. A video forensic technique for detecting frame deletion and insertion. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6226-6230). IEEE.

[30] Group of pictures. https://av-info.eu/index.html?https&&&av-info.eu/video/GOP.html. Último acceso 10/06/2022.

[31] ¿Qué es el formato H.264? Irene. Febrero 2012. https://www.leawo.com/es/knowledge/h264.html. Último acceso 17/06/0222.

[32] Ochoa-Domínguez, H.de J., Mireles-García, J., & Cota-Ruíz, J. de D.. 2007. Descripción del nuevo estándar de video H.264 y comparación de su eficiencia de codificación con otros estándares. Ingeniería, investigación y tecnología, 8(3), 157-180. Recuperado en 20 de diciembre de 2021, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-77432007000300004&lng=es&tlng=es. Último acceso 17/06/2022.

[33] Tayseer, M., Mohammad, J., Ababneh, M., Al-Zoube, A., & Elhassan, A. 2020, April. Digital Forensics and Analysis of Deepfake Videos. In 11th International Conference on Information and Communication Systems (ICICS).

[34] Recurrent Neural Networks. IBM Cloud Education. 14 de septiembre 2020. https://www.ibm.com/cloud/learn/recurrent-neural-networks. Último acceso 17/06/2022.

[35] Min Lee,J. CS3750, University of Pittsburgh. Recurrent neural networks and Long-short term memory. https://people.cs.pitt.edu/~jlee/papers/cs3750_rnn_lstm_slides.pdf. Último acceso 17/06/2022.

[36] Orozco, A. S., González, D. A., Villalba, L. G., & Castro, J. C. H. 2012. Anomalías en el Seguimiento de Exif en el Análisis Forense de Metadatos de Imágenes de Moviles. Actas del XII Reunión Española sobre Criptología y Seguridad de la Información, Donostia-San Sebastián, España.

[37] Li, H., Chen, H., Li, B., & Tan, S. (2018, November). Can forensic detectors identify gan generated images?. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 722-727). IEEE.

[38] Redes neuronales. IBM Cloud Education. Agosto 2020. https://www.ibm.com/es-es/cloud/learn/neural-networks#toc-tipos-de-r-gVH015Od. Último acceso 10/06/2022.

[39] Que son las redes neuronales y sus funciones. Industria 4.0 Inteligencia Artificial. Octubre 2019. https://www.atriainnovation.com/que-son-las-redes-neuronales-y-sus-funciones. Último acceso 10/06/2022.

[40] O'Shea, K., & Nash, R. 2015. An introduction to convolutional neural networks.arXiv preprint arXiv:1511.08458.

[41] D. Güera and E. J. Delp, Deepfake Video Detection Using Recurrent Neural Networks. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163.

[42] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[43] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. 2018, December. MesoNet: a compact facial video forgery detection network. In 2018 IEEE international workshop on information forensics and security (WIFS) (pp. 1-7). IEEE.

[44] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[45] Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. September 2019. The State of Deepfakes: Landscape, Threats, and Impact. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf. Último acceso 17/06/2022.

[46] Long Short-term Memory RNN. 2021. Vennerød, C. B., Kjærran, A., & Bugge, E. S . arXiv preprint arXiv:2105.06756.

[47] VIOLA, Paul; JONES, Michael. Rapid object detection using a boosted cascade of simple features. En Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. Ieee, 2001. p. I-I.

[48] RÖSSLER, Andreas, et al.2018. Faceforensics: A large-scale video dataset for forgery detection in human faces.arXiv preprint arXiv:1803.09179, 2018.

[49] Xia, Zhiming, Tong Qiao, Ming Xu, Xiaoshuai Wu, Li Han, and Yunzhi Chen. 2022. Deepfake Video Detection Based on MesoNet with Preprocessing Module. Symmetry 14, no. 5: 939. https://doi.org/10.3390/sym14050939. Último acceso 17/06/2022.

[50] Furht, Borko. (1999). Image presentation and compression.

[51] Sit, Muhammed & Demiray, Bekir & Xiang, Zhongrun & Ewing, Gregory & Sermet, Yusuf & Demir, Ibrahim. 2020. A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources. 10.31223/osf.io/xs36g.

[52] KINGMA, Diederik P.; BA, Jimmy. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[53] Xia, Zhiming and Qiao, Tong and Xu, Ming and Wu, Xiaoshuai and Han, Li and Chen, Yunzhi. *Symmetry* 2022,*14*(5), 939. Deepfake Video Detection Based on MesoNet with Preprocessing Module. https://doi.org/10.3390/sym14050939. Último acceso 14/06/2022.