Propuesta de Trabajo Final de Maestría 2024

Título tentativo: Detección de suplantacion de identidad usando tecnicas de reconocimiento facial con Machine Learning.

Alumno: Pedro Rostagno

Tutor: Daniel Acevedo

1. Introducción

1.1. Contexto

Según un informe reciente, más del 57% de los bancos y Fintech en EEUU y Reino Unido pierden más de USD 500.000 por año debido al fraude [1] y otro informe establece perdidas en el orden de los 23 mil millones de dólares totales debido a fraude por suplantación de identidad [2]. Estos fraudes donde tradicionalmente se usa ingeniería social para engañar a la víctima, están empezando a contar con mejores tecnologías gracias a la inteligencia artificial y la aparición de los DeepFakes (uso de IA para generar un rostro falso y potencialmente una suplantación de identidad), por lo que tener herramientas para detectar si un rostro es falso o no se vuelve un paso critico en la seguridad.

Zhang et al., 2012, presenta una base de datos compuesta de 50 sujetos genuinos y a partir de estos se crean los rostros falsos en tres calidades diferentes (alta, media y baja). Los tipos de ataques que se utilizan son fotos impresas deformadas, fotos cortadas y videos reproducidos. Se explora el uso de la información de alta frecuencia para detectar fraudes faciales, utilizando Support Vector Machine (SVM) y Difference of Gaussians (DoG) para distinguir entre rostros verdaderos y falsos.

Yu et al., 2022, proporciona una revisión sobre el uso de distintas técnicas de Deep Learning para la detección de fraudes faciales (face anti-spoofing, FAS), una tecnica para proteger los sistemas de reconocimiento facial contra los llamados ataques de presentacion (presentation attacks, PAs). A medida que los ataques de suplantacion de identidad se vuelven mas sofisticados, las tecnicas tradicionales, como parpadeo de los ojos, movimiento del rosto, brillo en el rostro, entre otros, deben ir mejorando, ya que estos necesitan de videos interactivos de larga duracion, lo que los hace inconvenientes en la practica. En cambio, los metodos basados en Deep Learning demostraron un buen rendimiento. Algunas de las tecnicas que se mencionan son el uso de redes neuronales convolucionales (CNN), y redes neuronales recurrentes (RNN). Tambien se menciona el uso de mapas de profundidad para poder detectar detalles mas sutiles entre los rostros autenticos y falsos. Se destaca la importancia de diseñar modelos capaces de generalizar los ataques desconocidos.

Tolosana et al., 2020, realiza un analisis de las diferentes tecnicas de manipulacion facial, incluyendo los DeepFakes y metodos para detectarlos. Se destacan cuatro tecnicas principales para la manipulacion facial:

1. **Sintesis completa del rostro**: esta manipulacion genera imágenes de rostros completas, usualmente utilizando Generative Adversarial Networks (GAN), como StyleGAN. Se discuten metodos para detectar los PAs mencionando 'huellas digitales' introducidas por las GAN.
2. **Intercambio de identidad**: consiste en reemplazar en un video la cara de una persona por otra.
3. **Manipulacion de atributos**: este tipo de manipulacion se refiere a la edicion facial, donde se modifican atributos como el color del pelo o la piel, el genero, la edad, etc. En esta manipulacion se utilizan generalmente GANs como StarGAN.
4. **Intercambio de expresiones**: consiste en la manipulacion de las expresiones faciales de la persona, donde tambien se utilizan arquitecturas GANs, con tecnicas como Face2Face y NeuralTextures, donde se reemplaza las expresiones faciales de una persona en un video, por las

expresiones de otra persona. Esta tecnica es usada en videos donde se muestra a alguien diciendo algo que nunca dijo.

En este ultimo articulo tambien se mencionan los retos para la deteccion de este tipo de fraudes, y a la presencia de ataques cada vez mas realistas y dificiles de detectar. Por ejemplo, en la manipulacion de Face Synthesis, se menciona que las arquitecturas GAN generan imágenes muy realistas, pero que estas dejan huellas que son faciles de detectar, pero que pasaria si es posible remover estas huellas o agregar algo de ruido a la imagen sin perder el realismo. Se destaca el uso de CNN para la deteccion de rostros falsos.

1.2. Problema

El problema que enfrenta la tecnología de reconocimiento facial es como se puede asegurar la autenticidad de un rostro, discriminando entre rostros genuinos y falsos. Este tipo de fraude puede tomar diversas formas y se está volviendo cada vez más complejo con el avance de la tecnología. En mi tesis busco desarrollar un sistema de machine learning capaz de detectar si el rostro que se presenta es autentico o falso.

1.3. Objetivo

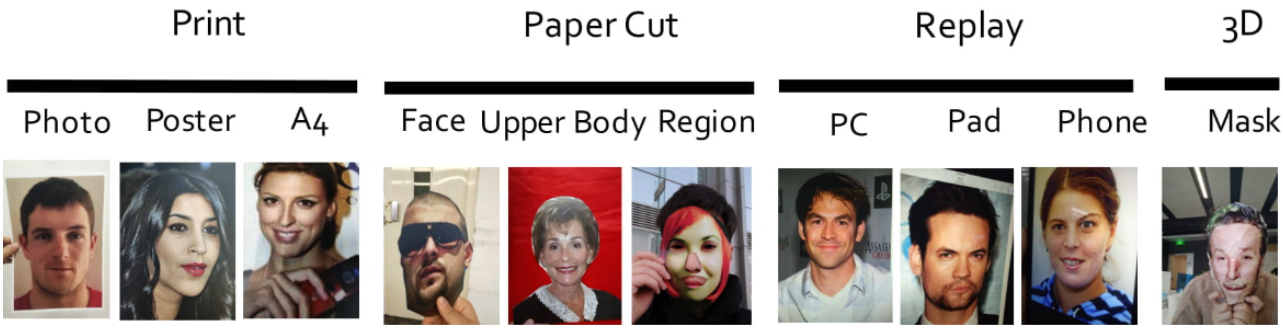
¿Es posible desarrollar un modelo de deteccion de suplantacion de identidad que mantenga un nivel de precision alto y que este optimizado para correr en un dispositivo movil?

Criterio de éxito: lograr un modelo que pueda correr en un dispositivo movil de ultima generacion sin tener perdidas significativas en la precision del modelo.

2. Datos

El Dataset a utilizar es un dataset público de uso no comercial, **'CelebA-Spoof'** [3]. Esta armado para la investigación en 'anti-spoofing' facial, es decir, para detectar si un rostro es real o falso. Este dataset está construido sobre un dataset llamado 'CelebA' (Celebrity Faces Attributes) [4] que contiene más de 200 mil imágenes de celebridades, y es utilizado para entrenar modelos de computer visión en tareas como reconocimiento facial y detección de rostro.

Se reporta que **'CelebA-Spoof'** esta compuesto por 625.537 imágenes de 10.177 sujetos y es uno de los datasets más grandes disponibles para investigación anti-spoofing. El dataset se divide en sets de entrenamiento y testeo, organizado por IDs de cada sujeto. Dentro de cada carpeta ID hay una subcarpeta 'live' que contiene imágenes reales de la persona, y otra subcarpeta 'spoof' con imágenes falsas generadas con distintas técnicas de suplantación de identidad. Estas técnicas incluyen fotografías impresas, fotografías impresas y recortadas, reproducciones en dispositivos móviles (teléfonos o tablets) y mascaras 3D.



Ejemplos de las tecnicas de spoof existentes en el dataset

El dataset posee anotaciones con 43 atributos, donde se incluyen atributos faciales (genero, edad, color de pelo, etc) heredados del dataset original 'CelebA', y atributos propios del dataset como tipo de 'spoof', condiciones de iluminación y ambiente. Los 40 atributos existentes, heredados del dataset original son: 5 o'clock shadow, arched eyebrows, attractive, bags under eyes, bald, bangs, big lips, big nose, black hair, blond hair, blurry, brown hair, bushy eyebrows, chubby, double chin, eyeglasses, goatee, gray hair, heavy makeup, high cheekbones, male, mouth slightly open, mustache, narrow eyes, no beard, oval face, pale skin, pointy nose, receding hairline, rosy cheeks, sideburns, smiling, straight hair, wavy hair, wearing earrings, wearing hat, wearing lipstick, wearing necklace, wearing necktie, young. Y los tres 'nuevos' atributos del dataset 'CelebA-Spoof' son: spoof type, illumination condition, environment.

Ejemplo de 4 sujetos que comparten 6 atributos aleatorios (male, big nose, young, eyeglasses, smiling, narrow eyes)

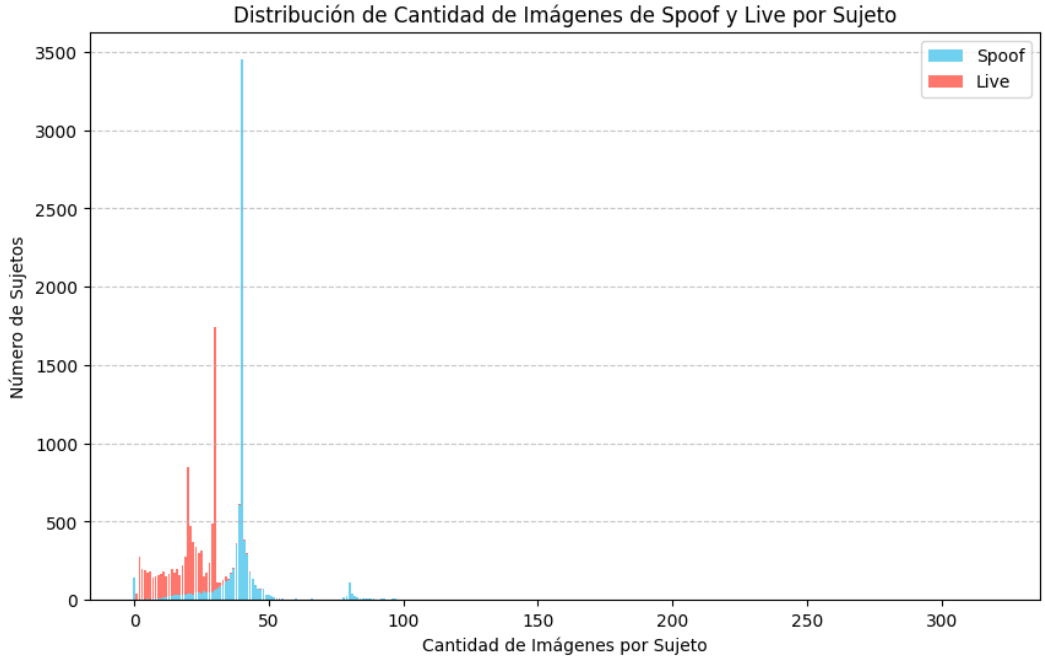
Ejemplo de 4 sujetos que comparten 5 atributos aleatorios (male, receding hairline, mustache, black hair, bags under eyes)

El dataset podría contener varios sesgos inherentes debido a la composición de las imágenes. Uno de los posibles sesgos está relacionado al género, ya que el 42% de los sujetos son masculinos en el dataset de entrenamiento y 38% en el dataset de testeo, lo que implica una subrepresentación de este grupo en el dataset. Esto podría influir en el rendimiento del modelo, ya que al estar entrenado en su mayoría con ejemplos femeninos es posible que tenga menos precisión al identificar rostros masculinos. Otro punto importante es que no existe un atributo relacionado a la raza o el color de piel, por lo que no se puede saber si existe un grupo étnico que este subrepresentado, lo que podría derivar en un mal rendimiento si se quiere detectar un spoof en una persona perteneciente a una raza subrepresentada en el entrenamiento del modelo. Es importante evaluar los resultados del modelo en distintos grupos para garantizar que el modelo funcione equitativamente para los distintos grupos (ya sea diferencias de género, raza, color de pelo, etc.).

El dataset se encuentra separado en dos carpetas, test y train. Y la cantidad de imágenes es la siguiente:

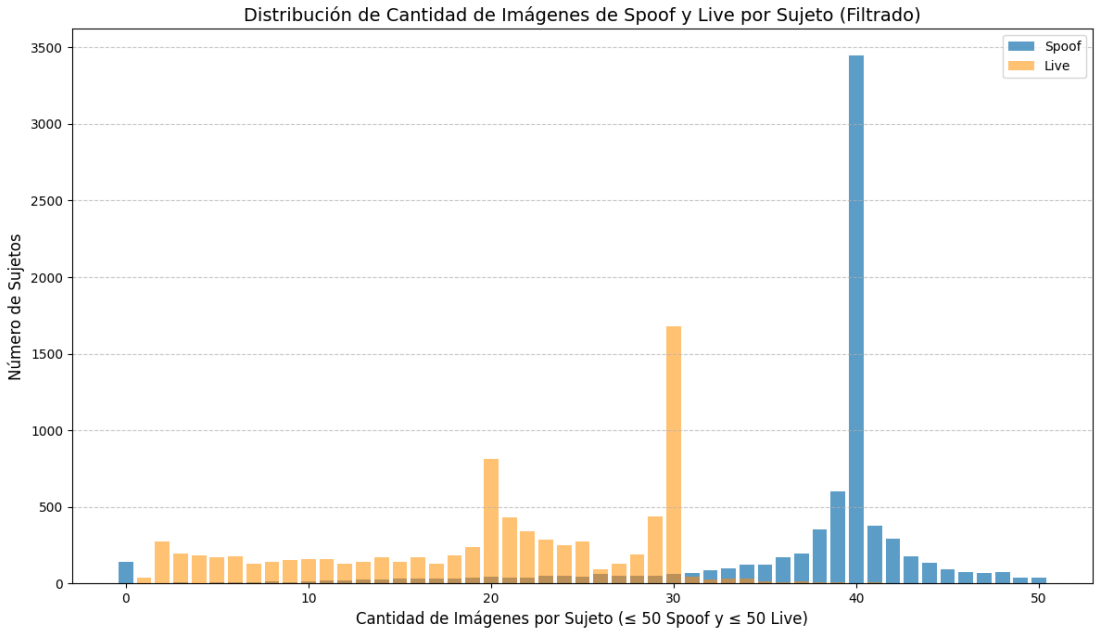
	Train	**Test**	**Suma**
Live	164.484	19.923	184.407
Spoof	329.921	47.247	377.168
			561.575
Cantidad de sujetos	8.192	1.004	9.196

El analisis de la cantidad de imágenes no se condice con lo reportado en el paper, ya que se reportan 625.537 imágenes de 10.177 sujetos, pero al revisar los datos, la cantidad de imágenes es de 561.575 y la cantidad de sujetos 9.196. Es importante analizar tambien la distribucion de la cantidad de imágenes de los sujetos, para ver si existen sujetos con pocas imágenes y otros con muchas.



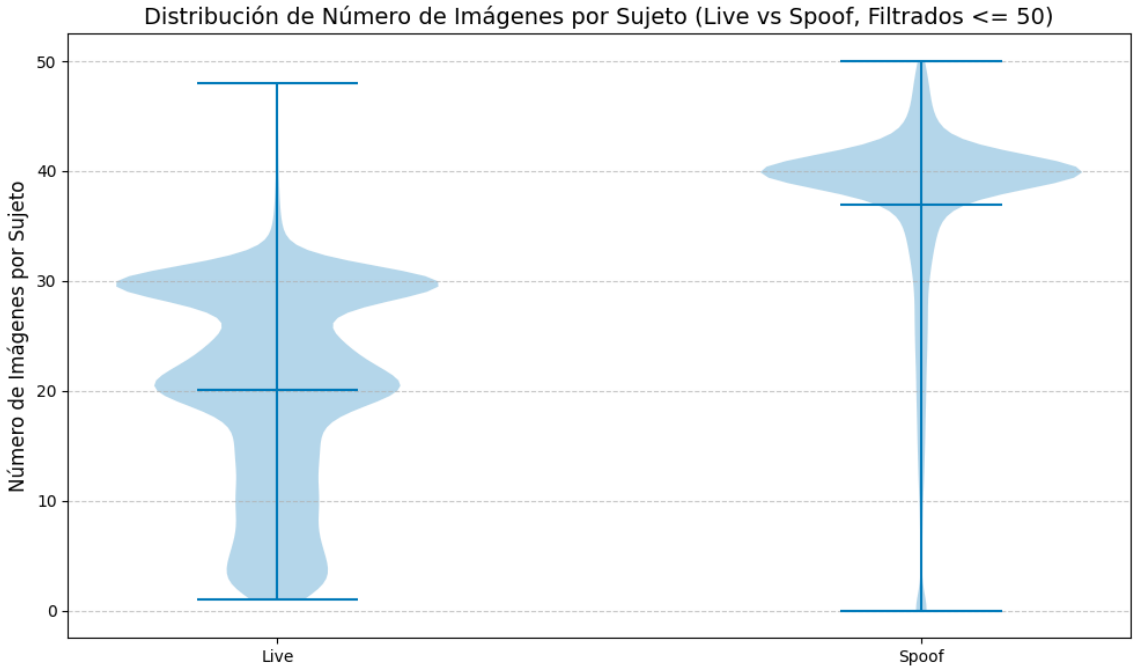
Distribucion de cantidad de imágenes por sujeto

A partir del gráfico anterior, se observa que la mayoría de los sujetos cuentan con entre 0 y 50 imágenes. Un análisis más detallado revela que hay 559 sujetos en la categoría spoof con más de 50 imágenes, mientras que solo 2 sujetos en la categoría de imágenes reales superan esa cantidad. Para lograr un modelo más balanceado y evitar el sobreentrenamiento en casos específicos, sería conveniente excluir del dataset a estos sujetos con un volumen excesivo de imágenes. A continuacion se muestra una distribucion con estos sujetos con grandes cantidades de imágenes filtrados.



Analisis para sujetos con cantidad de imágenes entre 0 y 50

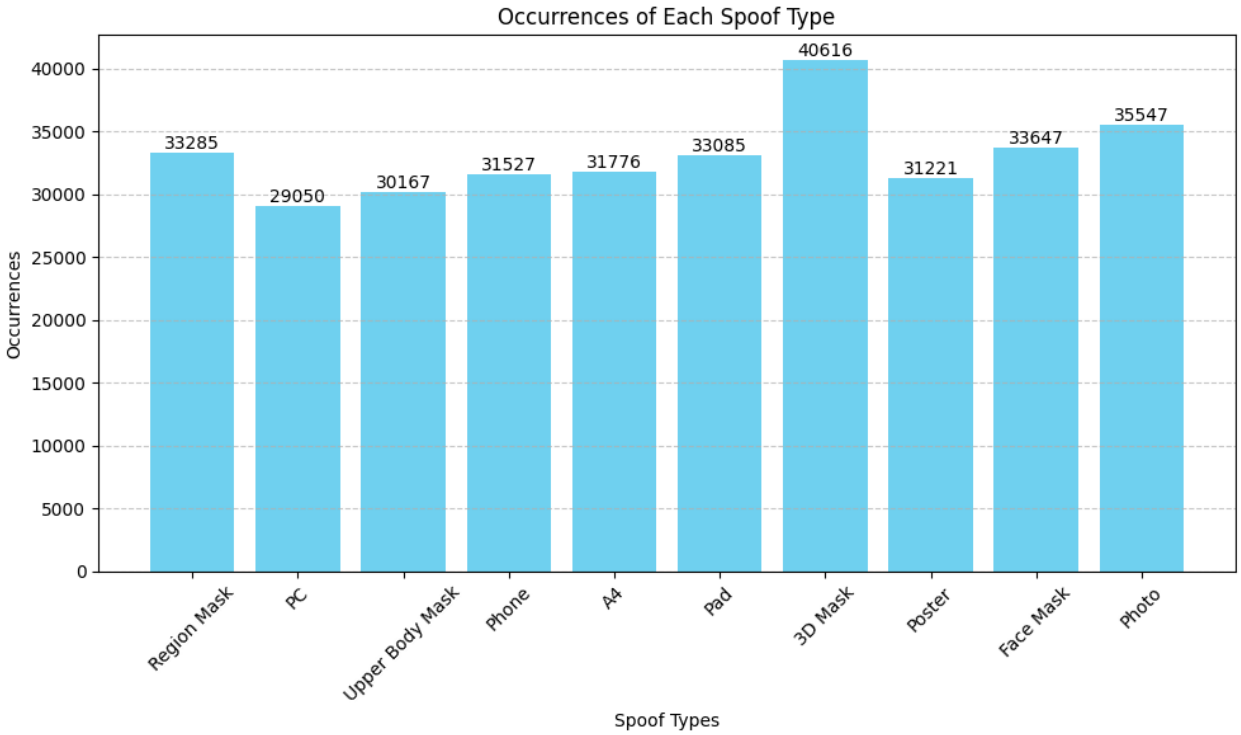
A partir del gráfico, se observa que existen sujetos sin imágenes spoof (un total de 143 sujetos), lo que indica una posible asimetría en la cantidad de datos entre las categorías. Además, el promedio de imágenes por sujeto es mayor en la categoría spoof en comparación con la categoría live, lo que sugiere una mayor variedad de datos en las muestras falsificadas. También se aprecia que la cantidad de sujetos con imágenes spoof alcanza un pico significativo alrededor de las 40 imágenes, mientras que en la categoría live el pico es más bajo y se distribuye de manera más uniforme entre 10 y 30 imágenes. Esto podría indicar una sobrerepresentación de imágenes spoof para ciertos sujetos, lo cual es importante tener en consideracion para evitar posibles sesgos en el entrenamiento del modelo. Esta distribución no balanceada sugiere la necesidad de un preprocesamiento para equilibrar ambas categorías y minimizar el riesgo de sobreajuste en el modelo, ya que el modelo podria aprender a clasificar incorrectamente, favoreciendo la categoria spoof, lo que podria llevar a que el modelo detecte bien los spoof pero no asi los casos reales, afectando la precision del modelo.



Distribucion de la cantidad de imágenes por sujeto en el dataset

En la imagen anterior podemos observar un grafico de violin donde se muestra la distribucion del numero de imágenes por sujeto, para las categorias Live y Spoof, filtrando para sujetos con menos de 50 imágenes. La categoria live presenta una distribucion mas ancha, lo que representa una mayor variabilidad en la cantidad de imágenes por sujeto, con dos picos en 20 y 30 (como se ve tambien en el grafico anterior del box plot), en cambio para los spoof, se ve una menor variabilidad en los numeros de imágenes y una mayor concentracion alrededor del pico de 40 imágenes por sujeto, muy cercano a su mediana.

Respecto a los diferentes tipos de spoof que se encuentran en el dataset de entrenamiento, se puede ver en el siguiente grafico que los datos estan relativamente balanceados y que no se ve una sobrerepresentacion por algun spoof especifico. El tipo de spoof con mayor representacion es la mascara 3D, con 40.616 imágenes y el de menor representacion es PC, con 29.050 imágenes. Al encontrarse balanceado el dataset, es beneficioso para el modelo a entrenar ya que se reduce la probabilidad de tener algun sesgo hacia algun tipo especifico de spoof.



Cantidad de imágenes según el tipo de spoof

3. Metodología

Preprocesamiento de los datos:

- Analisis y limpieza de los datos: revision de los datos y posible eliminacion de ciertas imágenes.
- Balanceo de clases: realizar un submuestreo para balancear las clases.
- Procesamiento de imagenes: analizar la calidad de las imágenes en condiciones de baja iluminacion y considerar la posibilidad de mejorar la misma.

Modelos:

- Voy a utilizar modelos de redes neuronales convolucionales como MobileNet para establecer un modelo base que me sirva como punto de referencia. Tambien planeo analizar el uso de modelos mas livianos que puedan correr en un celular.
- Analizar el uso de modelos pre-entrenados y realizar un fine tunning para adaptarlos al dataset y la deteccion de spoofing. Luego voy a comparar estos resultados contra el modelo base que voy a usar de referencia.
- Voy a explorar el uso de transformers, que en los papers leidos se menciona que tienen gran potencial para detectar suplantacion de identidad y este enfoque esta poco explorado.

Tecnologia:

- En pricipio los modelos van a ser implementados utilizando PyTorch y voy a analizar el uso de Lightning (tambien conocido como PyTorch Lightning) ya que permite estructurar mejor el codigo y crear proyectos mas complejos.
- Al necesitar procesar grandes cantidades de imágenes y no contar con una GPU fisica disponible, se va a utilizar algun proveedor en la nube para el procesamiento de los datos. Algunas opciones son: Google Colab, Kaggle o Hugging Face.

Prevencion o deteccion de sesgos:

- Como el dataset no posee ningun atributo respecto a la etnia, voy a enriquecer los atributos utilizando algun modelo existente que me permita etiquetar por etnias, como por ejemplo FairFace.
- Realizar un analisis exploratorio utilizando los nuevos atributos del dataset para ver si existen etnias que esten subrepresentadas.
- Evaluar el rendimiento del modelo por subgrupos para comparar si existen diferencias en los resultados según las diferencias en genero o etnias.
- Si el modelo es vulnerable a ciertos subgrupos y presenta sesgos, analizar un reentrenamiento del modelo para mitigar este problema.

Originalidad:

- Como las CNN tradicionales ya fueron utilizadas en este dataset, voy a explorar utilizar un modelo hibrido de CNN-Transformers para mejorar la deteccion anti-spoofing ya que este camino no fue explorado.
- Me gustaria desarrollar una version del modelo que pueda correr en dispositivos moviles.
- En el caso de encontrar resultados interesantes respecto a un modelo que presenta sesgos hacia ciertos subgrupos, voy a realizar un analisis detallado del mismo.

4. Cronograma preliminar

Tarea / Mes	**1**	**2**	**3**	**4**	**5**	**6**
Analisis y preparacion de los datos	X					
Desarrollo de modelo base de referencia		X				
Diseño y desarrollo de modelo con transformers		X	X			
Diseño y desarrollo de modelo hibrido (CNN-Transformer)			X	X		
Optimizacion de hiperparametros				X		
Analisis de sesgos y validacion				X	X	
Documentacion	X	X	X	X	X	X

Bibliografía inicial

Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., & Li, S. Z. (2012, March). A face antispoofing database with diverse attacks. In 2012 5th IAPR international conference on Biometrics (ICB) (pp. 26-31). IEEE.

Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., & Zhao, G. (2022). Deep learning for face anti-spoofing: A survey. IEEE transactions on pattern analysis and machine intelligence, 45(5), 5609-5631.

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.

Referencias

[1] https://www.alloy.com/state-of-fraud-benchmark-report-2024#component-marketo-embed

[2] https://www.aarp.org/money/scams-fraud/info-2024/identity-fraud-report.html

[3] https://github.com/ZhangYuanhan-AI/CelebA-Spoof

[4] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html