Sistema de Prevención de Deepfake en Características Biométricas de Vídeos (noviembre 2024)

Yenifer Andrea Viteri Riascos y Miguel Felipe Cifuentes
Universidad de los Andes

***Resumen*—Este proyecto propone desarrollar un sistema de detección y prevención de deepfakes en tiempo real para videollamadas, utilizando técnicas avanzadas de análisis de patrones y características biométricas, el cual busca abordar el creciente desafío que representa la manipulación de videos, conocida como media sintética, que pone en riesgo la autenticidad de los participantes, la seguridad de las empresas y la confianza en las comunicaciones virtuales.**

Con la implementación de este sistema, se garantizará la integridad de la información, la protección de los datos personales y el fortalecimiento de la seguridad en las interacciones digitales. Además, se evaluará su efectividad en escenarios prácticos y se desarrollarán medidas innovadoras contra posibles manipulaciones, este sistema podrá integrarse con diversas herramientas corporativas, permitiendo una mayor visualización y monitoreo de los funcionarios y aplicaciones utilizadas, lo que resultará en un control más robusto y efectivo a nivel general.

I. Introducción

En la actualidad, existen numerosas aplicaciones de inteligencia artificial que han simplificado significativamente el proceso de simular a una persona, logrando niveles de precisión tan altos que pueden ser indistinguibles de imágenes o videos reales. Estas tecnologías son aprovechadas por muchas empresas para mejorar la atención al cliente a través de portales web o medios digitales, ofreciendo experiencias más personalizadas y precisas, sin perder el toque "humano"[1].

Sin embargo, estas mismas aplicaciones están generando un riesgo creciente para las organizaciones. Su capacidad para replicar con exactitud la identidad de empleados, incluyendo aquellos en cargos altos con altos niveles de autorización, está siendo explotada para realizar suplantaciones que permiten ejecutar directrices de forma fraudulenta. Esto ha resultado en pérdidas económicas significativas para muchas empresas[2].

A pesar de este riesgo emergente, la mayoría de las organizaciones carecen de controles adecuados para enfrentarlo, dejando a sus sistemas vulnerables frente a estas amenazas, que continúan evolucionando y poniendo en peligro tanto la seguridad como la reputación corporativa. Es fundamental abordar este desafío con herramientas y estrategias diseñadas específicamente para mitigar el impacto de estas tecnologías en el entorno empresarial.

Con base en lo anterior, se ha diseñado una herramienta innovadora que permitirá detectar y prevenir, en tiempo real, ataques de deepfake durante videollamadas. Esta solución informará de manera inmediata a los participantes de la reunión sobre posibles manipulaciones, enviando también alertas a los sistemas de monitoreo de la empresa. Esto no solo fortalecerá la capacidad de respuesta ante estos ataques, sino que también aportará valor agregado al ecosistema de sistemas de información, permitiendo una defensa integral frente a posibles intentos de suplantación de identidad.

En caso de que un intento de ataque durante una videollamada no tenga éxito, los atacantes podrían recurrir a otros aplicativos o medios para comprometer a un funcionario. Gracias a esta herramienta, la organización podrá generar alertas oportunas y establecer controles adicionales para mitigar estos riesgos, mejorando la seguridad global[3].

La herramienta está diseñada para ser de aprendizaje continuo, lo que le permitirá adaptarse constantemente a nuevos patrones y amenazas. Además, su integración con los diferentes sistemas de información corporativos será ágil y sencilla mediante el uso de APIs, facilitando su incorporación con los validadores de identidad existentes y maximizando su efectividad en el entorno empresarial.

II. SOLUCIÓN Y REQUERIMIENTOS

Este proyecto propone un detector de Deepfake en tiempo real durante videollamadas, utilizando técnicas avanzadas de análisis de patrones para identificar y prevenir transmisión de vídeos manipulaos, proporcionando una capa adicional de seguridad que permita garantizar la autenticidad de los participantes, mejorando significativamente la integridad y fiabilidad de las comunicaciones virtuales.

1) *Funcionales*
   a. Capturar el vídeo en vivo.

[1] Chesney, R., & Citron, D. (2019). *Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics*. Foreign Affairs.

[2] Heather Chen and Kathleen Magramo, CNN, February 2024, https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

[3] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Niessner, Octubre 2019, FaceForensics++: Learning to Detect Manipulated Facial Images.

b. Identificar la cantidad de rostros presentes dentro del vídeo.
c. Preprocesamiento de fotogramas (redimensionar, normalizar), se debe asegurar que todos los fotogramas del video estén en un formato uniforme para que el modelo pueda procesarlos de manera eficiente
d. Extraer características de los fotogramas, se identificarán patrones clave que puedan diferenciar un video real de uno manipulado.
e. Realización de la predicción (real vs. manipulado), deberá determinar si un fotograma es auténtico o ha sido manipulado mediante algoritmos de deepfake.
f. Visualización de los resultados en tiempo real, en este parte habrá una retroalimentación inmediata a los usuarios para alertar sobre posibles manipulaciones.

2) No funcionales

a. El prototipo debe utilizar algoritmos de cifrado como AES256, SHA256, RSA o cualquier otro que garantice la confidencialidad e integridad de la información en su transmisión.
b. El sistema debe ser capaz de manejar volúmenes altos de solicitudes de forma concurrente que no afecte el proceso de verificación ni generar una denegación de servicio (DDoS).
c. El sistema debe mantener una disponibilidad del 99,9%, asegurando su funcionamiento continuo y disponibilidad, para que pueda identificar en todo momento un deepfake.
d. Optimización del sistema para reducir el tiempo de inferencia.

III. DISEÑO

Un director financiero en Hong Kong fue víctima de fraude, el ataque comenzó con la suplantación del correo electrónico del funcionario, desde el cual se convocó a una reunión urgente con n grupo de empleados, durante la videollamada los atacantes utilizaron un software avanzado de inteligencia artificial para replicar la apariencia y voz del CEO de la compañía[4].

La representación fue tan convincente que los participantes en la reunión creyeron estar interactuando con el auténtico CEO, aprovechando esta confianza, el falso 'CEO' solicitó la transferencia urgente de 35 millones de dólares a cuentas en el extranjero, bajo el pretexto de cerrar una transacción comercial importante.

El funcionario responsable, confiando en la autenticidad visual y en la jerarquía del solicitante realizó la transacción, una vez realizado el envío, los ciberdelincuentes dispersaron rápidamente los fondos a múltiples cuentas, dificultando su rastreo y recuperación.

Con la implementación de la herramienta de prevención de deepfake, será posible detectar en tiempo real si algún participante de una reunión utiliza inteligencia artificial para simular la identidad de otra persona.

El funcionamiento de la herramienta comienza al recopilar datos preliminares del participante, como la dirección IP, geolocalización e ID del dispositivo, esta información es útil para identificar posibles atacantes. Antes de ingresar a la reunión, se solicita al usuario su consentimiento para el tratamiento de datos personales, ya que se recopilarán datos biométricos, si el participante no autoriza el uso de esta información, se genera una alerta automática para el área de antifraude o el departamento designado por la empresa, que monitoreará al usuario de forma activa.

En caso de que el usuario acepte, podrá ingresar a la reunión, pero la herramienta activará su proceso de validación, este proceso analiza el video en tiempo real, segmentándolo en múltiples frames que son evaluados utilizando datasets especializados y las políticas internas de la empresa.

El análisis genera un resultado que, en caso de detectar irregularidades, activa una alerta inmediata para todos los participantes de la reunión. Además, envía notificaciones automáticas a los sistemas de seguridad de la compañía, como el SOC, SIEM u otras plataformas de monitoreo disponibles, esto permite que las áreas responsables implementen medidas correctivas y controles adicionales, fortaleciendo la seguridad y mejorando continuamente la eficacia del sistema de prevención de deepfakes.

IV. ARQUITECTURA

Las herramientas basadas en machine learning requieren una fuente considerable de datos para ser entrenadas y alcanzar un alto nivel de eficiencia en la tarea asignada mediante múltiples iteraciones. Por esta razón, es crucial definir los datasets que se utilizarán para alimentar la Red Neuronal Convolucional (CNN) destinada a la identificación de posibles casos de spoofing-deepfakes, antes de proceder a detallar el análisis que se realizará.

Para lo cual se han seleccionado tres datasets principales que servirán como base, los dos primeros son repositorios ampliamente utilizados por diversas organizaciones, elegidos por la calidad y cantidad de datos que ofrecen, el tercero, en cambio, es de naturaleza dinámica, ya que puede variar según el cliente; sin embargo, los requisitos mínimos para este último son consistentes y se detallarán a continuación

[4] Heather Chen and Kathleen Magramo, CNN, February 2024, https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

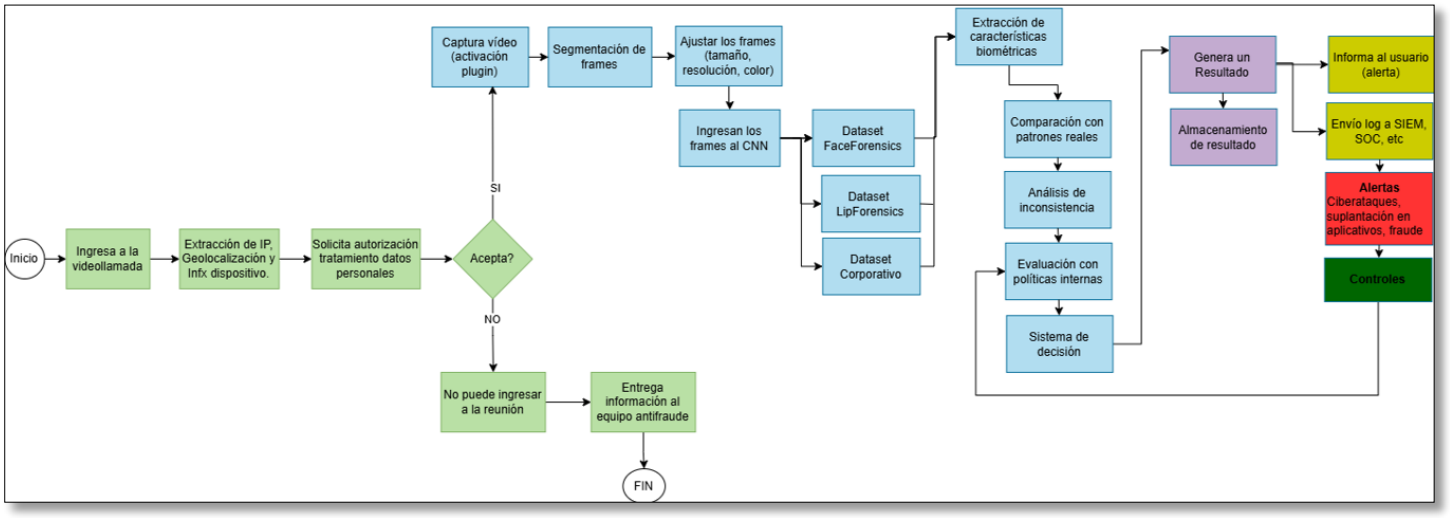

Ilustración 1. Caso de uso de herramienta de prevención

1) Primer Dataset FaceForensics++

Es un dataset especializado en suplantaciones faciales, desarrollado por investigadores de la Universidad de Cornell. Su propósito es entrenar modelos de aprendizaje profundo utilizando más de tres mil ejemplos de deepfakes, millones de imágenes y más de mil videos provenientes de diversas fuentes. Este dataset incorpora nuevos métodos de manipulación facial, como DeepFake, FaceSwap y texturas naturales, convirtiéndolo en una fuente valiosa de datos para redes neuronales convolucionales (CNN). Además, su implementación es computacionalmente ligera, lo que permite un uso más eficiente de los recursos del computador.

Cabe destacar que FaceForensics++ no solo ofrece una gran riqueza informativa, sino que también integra otros datasets menos sofisticados, como FaceSwap y Face2Face, lo que contribuye a crear una base sólida para la detección de manipulación facial[5].

2) Segundo DataSet LipForensics

LipForensics es un dataset que adopta un enfoque único para la detección de videos manipulados, enfocándose en irregularidades muy detalladas en el movimiento de los labios, para ello, utiliza una red espaciotemporal entrenada para reconocer los movimientos virtuales de la boca, también conocido como lectura labios, lo que le permite identificar con precisión el movimiento natural de los labios.

Además, los datos están clasificados como verdaderos o falsos, lo que facilita una rápida integración con otros datasets, al ser de código abierto, LipForensics está disponible para descarga en GitHub, lo que lo hace accesible y fácil de implementar en proyectos de detección de videos manipulados[6].

3) Tercer DataSet Datos del Cliente

Finalmente, el último dataset está compuesto por una serie de datos del usuario final proporcionados por la organización, dado que estos datos son altamente sensibles, es crucial garantizar su anonimizarlos para protegerlos y cumplir con las normativas legales vigentes, tema que se abordará al final de este apartado.

Este dataset es de gran importancia para la solución, ya que los dos anteriores proporcionan una base sólida para la identificación, mientras que este agrega valor al utilizar los datos específicos de los usuarios finales de la organización donde se desplegará la herramienta.

Su diseño se compone de dos partes, en primer lugar, el usuario final deberá grabarse leyendo prosa del autor nicaragüense Rubén Darío. El objetivo es capturar las variaciones en la voz del usuario, como entonación, acentuación y ritmo, para ello, se han seleccionado obras literarias con prosa que ofrezcan buena puntuación y un uso del idioma local, preferiblemente con una métrica regular y un número elevado de sílabas, sin resultar agotadoras para el lector. Por ejemplo, el poema 'Divagación' tiene una métrica constante de 12 sílabas, con una correcta puntuación y musicalidad, lo que permite al sistema registrar cambios en la entonación.

Además, esta tarea resulta más amena para el usuario final, ya que requiere concentración para una lectura adecuada, sin presentar mayor dificultad.

El ejercicio consistirá en lecturas secuenciadas de 1 minuto, evitando el cansancio del lector y generando un promedio de 5 a 7 minutos de audio, durante la lectura, se grabarán los movimientos de los labios y el rostro del usuario para complementar los otros dos datasets.

Por otro lado, este tipo de datos deben tener otro tipo de requerimientos para salvaguardar los datos. Como ya ha sido estipulado anteriormente, los datos una vez tomados no pueden ser alterados ni modificados de ninguna manera, el derecho a ser olvidado por parte del usuario tiene que ser honorado según la ley colombiana vigente. Por otro lado, se recomienda a la organización que use un tipo de cifrado fuerte con algoritmos como SHA 256 para encriptar la data biométrica, al igual que implementar prácticas robustas de seguridad para el manejo de

[5] FaceForencics, https://github.com/ondyari/FaceForensics

[6] LipForencics, https://github.com/ahaliassos/LipForensics

las llaves criptográficas acordes a las políticas de seguridad de la empresa.

Se recomienda utilizar el servicio de Amazon S3 para almacenar los datos, integrándolo con AWS CloudHSM para maximizar la seguridad. Esta integración permite proteger las claves de cifrado utilizadas en los datos biométricos y restringir el acceso a personal autorizado únicamente. Además, es crucial incluir este almacenamiento y su configuración en las auditorías de seguridad de la empresa para garantizar el cumplimiento de las normativas y las mejores prácticas en protección de datos.

Adicionalmente, es importante usar mecanismos de seguridad como "Masking" que consiste en reemplazar partes sensitivas de los datos biométricos con caracteres aleatorios o tokens para reducir el riesgo de que estén expuestos. En caso de que sea posible, también se podría considerar usar plantillas basadas en representaciones matemáticas de los datos biométricos, en vez de los datos crudos en sí mismos.

Con estos mecanismos y acompañados de las políticas ya existentes de seguridad de la empresa, los datos biométricos se pueden guardar de manera segura reduciendo el riesgo a cambios, errores o pérdida de la información, es importante por último aclarar que estas no son una camisa de fuerza para la organización, sino por el contrario una ayuda para reforzar las políticas existentes.

a) Cumplimiento de Salvaguardado de Datos

En el contexto de este proyecto, no se prevé prestar servicio inicialmente a organizaciones o usuarios que no sean ciudadanos colombianos, por ello, el enfoque principal ha sido asegurar la adherencia a la legislación colombiana; sin embargo, también se han considerado normativas internacionales como el Reglamento General de Protección de Datos de la Unión Europea (GDPR) y la Ley de Privacidad del Consumidor de California (CCPA), para garantizar el cumplimiento de los requisitos mínimos que estas establecen.

De acuerdo con lo anterior, para este caso se contempla la ley 1581 de 2012 de Protección de Datos Personales, la cual establece que el titular del dato tiene el derecho de acceder, rectificar, modificar y eliminar sus datos personales, también define las obligaciones de quienes recopilan y procesan dicha información[7].

Por ello, la organización debe proporcionar su consentimiento explícito y firmado para la recolección y procesamiento de los datos del titular antes de realizar cualquier prueba. Esto garantiza que el usuario esté debidamente informado sobre el uso de sus datos y asegura el cumplimiento del Decreto 1377 de 2013. Además, se implementarán estrictos controles de seguridad para proteger los datos en tránsito, en uso y en almacenamiento, los cuales se detallarán en las secciones sobre el despliegue de la solución.

En caso de que el titular del dato ya no esté asociado con la organización, se mantendrán por un periodo de tiempo 3 meses, pero con la salvedad de que si el titular del dato solicita su eliminación se procederá con un borrado seguro de esta información.

Asimismo, en alineación con la Resolución 1369 de 2024, se integrará o se habilitará un canal atención al cliente para que puedan presentar quejas, dudas o reclamos relacionados con el manejo de sus datos y lo cuales están alineados con los requerimientos establecidos por la Superintendencia de Industria y Comercio (SIC).

A. Captura del "Stream" del vídeo de la videollamada

Para el correcto funcionamiento de la herramienta, es necesario capturar audio y video, que luego se transmitirá al servidor donde está alojada la red CNN, esta captura requiere que la herramienta también pueda transferir y recibir información desde el servidor, por eso se ha decidido diseñarla como una extensión para navegadores basados en Chromium, por dos motivos principales.

- El primero es la diversidad y popularidad de estos navegadores, como Vivaldi, Chrome, Opera y Edge (basados en Chromium), entre otros.
- El segundo, al ser de código abierto, ofrecen una arquitectura común que facilita la flexibilidad de la extensión, además, permiten que la API interactúe de manera extendida y eficiente con el navegador y las páginas web.

B. Objetivos de la API:

- Acceso a WebRTC: la API tiene la facultad de acceder al micrófono y cámara del usuario para capturar el flujo de datos de audio y video en tiempo real.
- WebSockets: Es necesario que se establezca un protocolo de comunicación bidireccional entre el navegador y el servidor en la nube, permitiendo la transmisión en tiempo real; sin embargo, no se desarrollará un nuevo protocolo para este proyecto, ya que existen suficientes métodos y protocolos que garantizan una transmisión fiable y de alta calidad.
- Generación de alertas en tiempo real: En función de la respuesta del servidor (positiva o negativa), la herramienta debe notificar al usuario de manera eficiente y rápida.

La herramienta está diseñada para funcionar durante videoconferencias sin importar la plataforma, como Google Meet, Zoom o Microsoft Teams, la única excepción serían plataformas cerradas como WebEx, ya que para que la herramienta funcione, la videoconferencia debe ser vía web y no a través de aplicaciones descargables.

[7] Ley de protección de datos 1581 de 2012, https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981

Al capturar audio y video desde aplicaciones de escritorio requeriría colaboración con las organizaciones propietarias (Meta, Google, Microsoft, Cisco, etc.), para no infringir derechos de autor o propiedad intelectual, además, las aplicaciones de escritorio utilizan diferentes puertos de red y características específicas que demandarían una adaptación de la API a cada caso.

C. Procesamiento de frame

Una vez capturado el video durante la videoconferencia, este es transmitido al servidor web, donde la red CNN analiza los datos de audio y video de manera secuencial e individual, esto significa que cada fotograma (frame) es procesado por separado.

El propósito de este enfoque es llevar a cabo un análisis detallado de los elementos más básicos, en este caso, los fotogramas. No se pretende establecer relaciones temporales entre ellos, ya que las manipulaciones de datos sintéticos podrían evadir este tipo de análisis al combinar elementos reales y alterados dentro de una misma transmisión.

El CNN extrae los fotogramas y los procesa mediante capas de análisis conocidas como "Pooling layers", cuyo objetivo es identificar y extraer las características más relevantes de cada uno, una vez completado este proceso, los datos extraídos se envían a las capas de conexión, donde se lleva a cabo la clasificación y predicción, es decir, el análisis final y la toma de decisiones.

D. Inferencias con el modelo de detección de Deepfake

Para el análisis y la posterior respuesta de la red convolucional, una vez separados los datos más relevantes, estos se comparan con los patrones previamente establecidos en los conjuntos de datos de entrenamiento (DataSets) descritos anteriormente, es importante aclarar que una CNN no realiza una comparación de datos en el sentido tradicional, sino que identifica patrones y relaciones basados en la información ya almacenada.

La ventaja de este enfoque es que cada nuevo fotograma procesado contribuye al aprendizaje y mejoramiento de los patrones existentes, en otras palabras, la red se encuentra en un estado constante de aprendizaje continuo, lo que incrementa su fiabilidad a medida que se nutre de más datos previamente alimentados.

E. Respuesta y notificaciones en tiempo real

Si el sistema detecta que un video está siendo manipulado, generará una alerta en la interfaz de la videollamada para advertir al usuario, para ello, se evaluarán los siguientes aspectos:

- Coincidencia de patrones: para identificar si los fotogramas han sido manipulados, se ha definido un índice de coincidencia alto, entre 0.90 y 1. Este rango ha sido seleccionado con el objetivo de minimizar el riesgo de falsos positivos y falsos negativos, garantizando así una mayor precisión en la detección.
- Las alertas que se generen notificarán a los usuarios de las videollamadas y a los sistemas de información que maneje la empresa como SOC, SIEM, etc, para generar posibles controles adicionales sobre un funcionario.

V. IMPLEMENTACIÓN

Para la implementación de la herramienta de prevención de deepfakes en tiempo real, es fundamental tener el aplicativo utilizado para las reuniones virtuales, dataset corporativo y considerar los sistemas de información y monitoreo existentes en la empresa.

El dataset corporativo servirá como una base inicial de rostros, permitiendo a la herramienta realizar un enrolamiento previo y certificado que garantice la identificación de los empleados de la empresa. Además, los sistemas de monitoreo, como el SOC o el SIEM, se integrarán para retroalimentar los reportes generados, facilitando la creación de alertas útiles y relevantes para la compañía.

La integración con nuestra herramienta se llevará a cabo a través de una API, lo que permitirá una conexión rápida y eficiente. Esto facilitará la personalización de la información requerida para interactuar con los diferentes sistemas, generando los logs y reportes necesarios para satisfacer las necesidades específicas de la empresa.

A continuación, se presenta la arquitectura técnica que garantizará el correcto funcionamiento y la efectividad de la herramienta.

Tipo Activo	Activo	Descripción
Hardware	Servidores	Servidor con GPU (NVIDIA A100)
	Almacenamiento	Amazon S3 20 TB
	Redes y Comunicaciones	Canal de internet mínimo 10 Gbps
Software	Sistema Operativo	Linux 2 (AWS)
	Servicios de videollamadas	Zoom, Microsoft Teams o Google Meet
	Base de Datos	Amazon RDS (PostgreSQL)
	Frameworks de Machine Learning	TensorFlow (Integrado al servicio de AWS)
	Middleware de Video	GStreamer (Integrado al servicio de AWS)
Nube	Balanceadores de carga	Elastic Load Balancer (ELB)
	Sistemas autenticación	AWS IAM (Integrado al servicio de AWS)

Tabla 1. Requerimientos técnicos

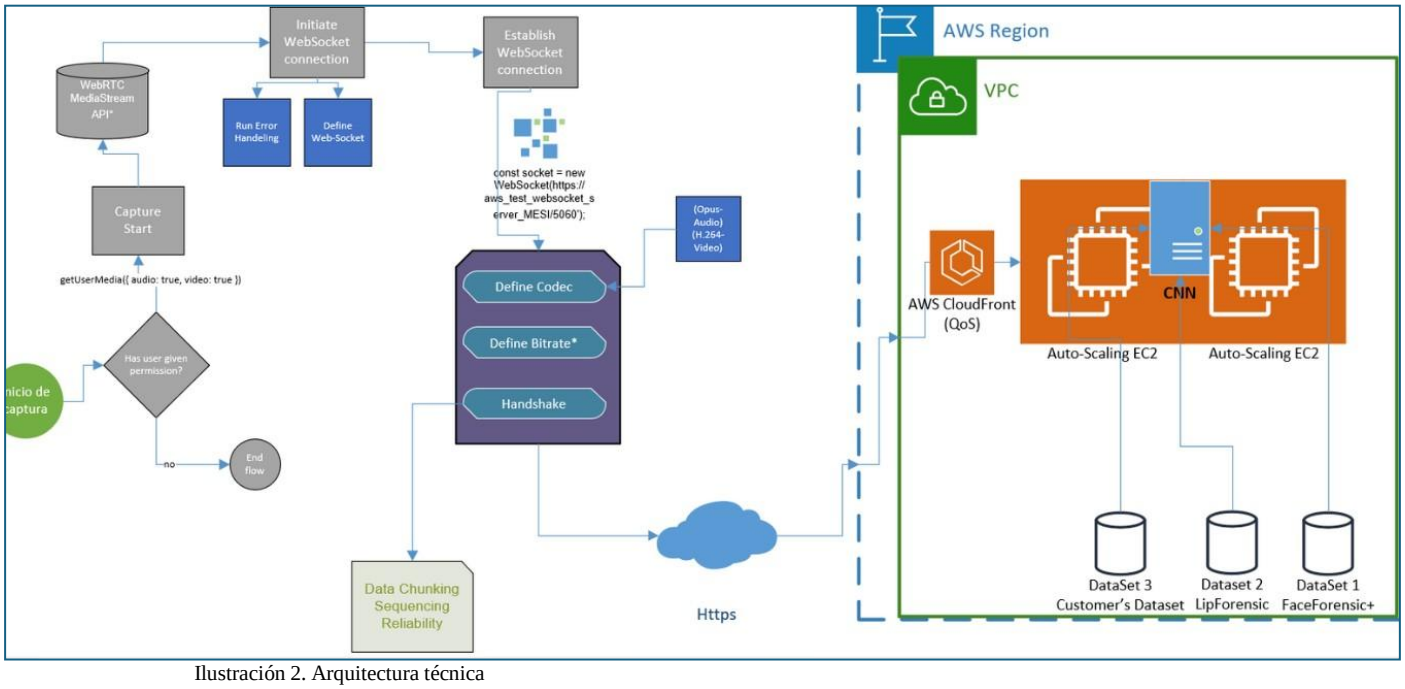

Ilustración 2. Arquitectura técnica

VI. DESPLIEGUE

Una vez completada la implementación, se realizarán pruebas con usuarios proporcionados por el cliente y posteriormente, se llevará a cabo un período inicial de aprendizaje, de 2 a 3 meses, durante el cual la herramienta recopilará datos sobre los comportamientos y rasgos biométricos de los funcionarios. Con esta información, será posible implementar las políticas de seguridad definidas por la empresa.

La información inicial proporcionada será clave para configurar de manera específica las alarmas, políticas y controles que la empresa considere necesarios para fortalecer su seguridad. Esto permitirá personalizar la herramienta según las necesidades y requerimientos particulares de la organización, asegurando un rendimiento óptimo y alineado con sus objetivos operativos.

Se entregará una documentación completa y detallada que incluirá instrucciones sobre el funcionamiento de la herramienta, así como guías paso a paso para su configuración e integración con los diferentes sistemas existentes en la empresa. Además, esta documentación contendrá ejemplos prácticos, mejores prácticas y recomendaciones para maximizar la eficacia de la herramienta en el entorno corporativo.

VII. METODOLOGÍA

En la ejecución de este proyecto hemos seleccionado la metodología en cascada debido a su enfoque estructurado y secuencial, ideal para proyectos con requisitos bien definidos desde el inicio. La metodología en cascada nos permite avanzar de forma ordenada a través de fases que tiene las cuales son: planeación, análisis de requerimientos, diseño, implementación, pruebas y despliegue, cada etapa debe completarse antes de pasar a la siguiente, asegurando la calidad de cada componente del proyecto antes de avanzar.

VIII. ANÁLISIS DE RESULTADOS

El proceso de obtención de resultados se llevó a cabo en varias fases, comenzando con el entrenamiento del modelo utilizando datasets de videos reales y falsos. El entrenamiento tuvo una duración de 8 a 9 horas por sesión, dado que los videos (de calidad mínima de 720p) se descomponen en miles de fotogramas para su análisis.

En la primera etapa, se utilizó un dataset de videos auténticos para establecer un punto de referencia confiable, asegurando que todos los videos fueran genuinos, ya que la inclusión de videos falsos en esta etapa podría comprometer el modelo. Posteriormente, se incorporó un segundo dataset compuesto por videos falsos, manteniendo la misma calidad de los originales para evitar fluctuaciones en los resultados.

El proceso incluyó técnicas como ajuste de tamaño, normalización, análisis de píxeles y clasificación. Los fotogramas se analizaron mediante filtros convolucionales 3D para capturar características espaciales y temporales, mientras que las capas de agrupamiento redujeron la complejidad computacional. Finalmente, las capas totalmente conectadas procesaron estas características, generando patrones numéricos que permitieron clasificar los videos como auténticos o manipulados.

Durante las 4 sesiones de entrenamiento realizadas, la precisión del modelo alcanzó un valor promedio de 0.95, aunque se observaron fluctuaciones en la validación debido a posibles casos de sobreajuste. Para mitigar esto, se optimizó el algoritmo y se ajustaron los datos de entrenamiento, enfocándose en patrones específicos, como rostros y audio.

Aunque la pérdida de entrenamiento disminuyó con cada sesión y la validación mejoró de forma constante, la detección de sobreajuste resalta la necesidad de incrementar la diversidad y

calidad de los datos, así como realizar más sesiones de entrenamiento en futuras iteraciones del modelo.

Para este prototipo teniendo en cuenta la limitante de energía y recursos, se realizaron 4 sesiones de entrenamiento, en donde se evaluaba al final la precisión del modelo a través del tiempo

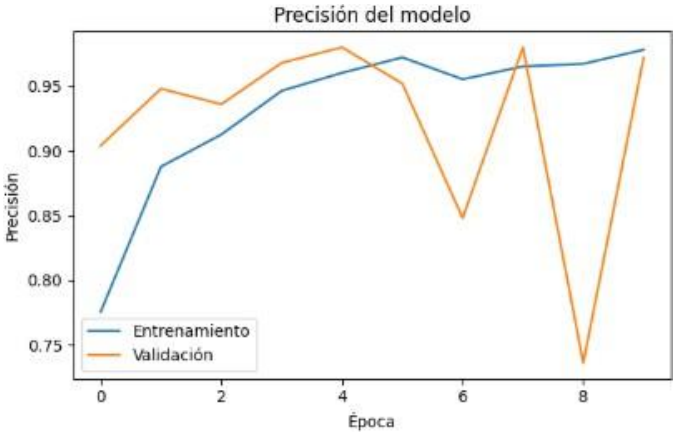

Ilustración 4. Precisión del modelo

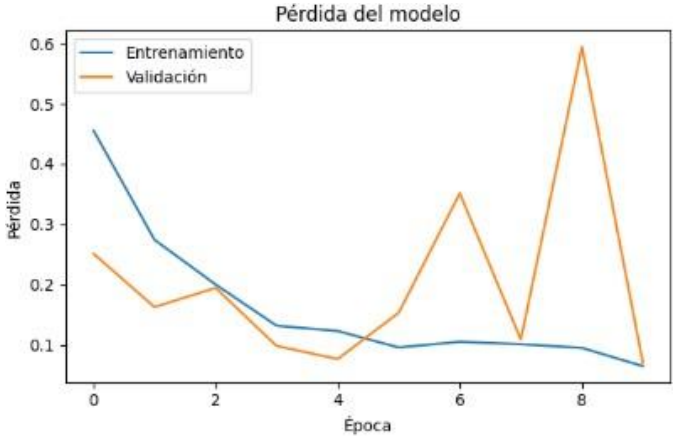

Ilustración 5. Pérdida del modelo

En la medida de los entrenamientos, la precisión del prototipo aumentaba con cada ronda alrededor de 0.95 lo que indicó que el entrenamiento se condujo de manera correcta con la posibilidad de aumentarlo usando videos de mayor calidad. La validación de la precisión también incrementó inicialmente, pero a medida que los entrenamientos fueron aumentando empezó a fluctuar lo que indicaba un posible caso de sobreajuste, por lo que se optimizó el algoritmo para enmendar esto al igual que buscar videos enfocados más hacia el rostro o el audio de manera que el modelo no solo memorizara los patrones sino también enfocarlo a que los identifique más eficazmente.

Por otro lado, la pérdida de entrenamiento disminuye con cada sesión, lo que sugiere que el modelo está aprendiendo y memorizando patrones de los datos. Aunque el desempeño en la validación es bajo al inicio, este mejora de forma constante con cada sesión, indicando que el modelo está generalizando mejor a los datos no vistos.

Sin embargo, se identifica una clara tendencia a la saturación del modelo, para esto es necesario aumentar la diversidad de los datos lo que en otras palabras significa, más datos y más sesiones de entrenamiento, que a futuro es un problema solucionable.

El prototipo utiliza un total de 1,650 neuronas, y su capacidad se incrementa proporcionalmente a los recursos de cómputo disponibles. Para garantizar precisión y validación, se decidió proceder conarro:

Indicador	Valor
Accuracy	0.9977
Precision	0.9980
Recall	0.9972
F1 Score	0.9976
AUC ROC	0.9999

Tabla 2. Métrica del modelo.

La métrica Accuracy representa las veces que el prototipo correctamente predijo el resultado. Precision indica cuando el prototipo predice un resultado positivo lo que nos indica el manejo de falsos positivos. Por otro lado, la métrica de Recall indica los casos reales positivos, lo que nos da el índice de Falsos negativos. El F1 Score es el índice armónico entre las métricas de Precision y Recall, un índice alto de esta métrica sugiere un buen balance. Y finalmente, el AUC ROC es el área inferior a la curva de ¨Receiver Operating Characteristic¨ lo que señaliza si el prototipo es bueno en distinguir entre casos positivos y negativos.

El bach size por el que se decidió fue de 16 con un intervalo entre fotogramas de 5, el tamaño mínimo de la cara para que sea aceptada fue de 64 pixeles, y el número máximo de fotogramas a procesar fue de 200, para que el prototipo pudiese procesar correctamente el video se requiere que haya un mínimo de rostros en el video, para que el video pudiese ser analizado es necesario que haya un mínimo de 7 fotogramas de un rostro que cumpla los parámetros anteriores. En sus etapas iniciales el modelo arrojó los siguientes resultados:

- Falsos positivos: en 160 instancias fueron incorrectamente clasificados como reales
- Falsos negativos: en 14 instancias fueron incorrectamente clasificados como falsos
- Verdaderos Positivos: en 5164 instancias fueron correctamente clasificados como reales
- Verdaderos Negativos: 4374 instancias fueron correctamente clasificados como falsos

Esto indica que el prototipo clasificó correctamente en un 96.4% de las veces los videos falsos, mientras que en las veces que predice que algo es verdadero acierta en un 99.7% de las veces. Por otro lado, el prototipo clasifica los datos correctamente en un 97.1% de las veces.

Para lograr el umbral deseado de un 99% es necesario más sesiones de entrenamiento junto con más datos para alimentarlo.

Se evidenció en algunos casos con ciertos videos falsos, el modelo los clasificaba como verdaderos con un índice de confianza del 87%, sin embargo, cuando esos videos que el prototipo catalogaba como verdaderos fueron presentados a otros modelos de detección estos también mostraban que el video era verdadero con un índice de confianza entre el 60% a 70% por lo que con más entrenamiento aplicando métodos adicionales para detección el índice puede ser reducido e inclusive correctamente clasificado. Es probable que los videos de mayor calidad superior o igual a 1080p60 con mayor movimiento sean clasificados como reales debido a las limitantes de procesamiento de fotogramas y constantes que tuvieron que ser adaptadas a la máquina donde se aplicaba el prototipo.

Por otro lado, las distribuciones de las instancias reales se encontraban centradas cerca al 0, con algunas instancias que tuvieron probabilidades ligeramente altas, lo que sugiere que el modelo tiene alta confianza al clasificar las instancias como reales.

Mientras que en las distribuciones para las instancias falsas está centrado cerca al 1, también con algunas instancias que presentaron probabilidades más bajas, lo que indica que el modelo de igual manera tiene un índice de confianza alto en clasificar las instancias como falsas.

Esto indica que el prototipo está correctamente calibrado; sin embargo, también puede sugerir que exista un exceso de confianza, lo cual podría ser resultado de un desequilibrio entre los datos reales y falsos. Para abordar esto, es necesario realizar más sesiones de entrenamiento y tener un mayor control sobre la cantidad y diversidad de videos presentados al prototipo, garantizando un balance equitativo.

REFERENCIAS

[1] B. A. y Arcas, D. (2022). Do large language models understand us? Daedalus, 151(2), 183-197.
[2] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 35(1), 53-65. https://doi.org/10.1109/MSP.2017.2765202
[3] https://www.researchgate.net/publication/373700317_Artificial_Neural_Networks_An_Overview
[4] Samoilenko, S.A., Suvorova, I. (2023). Artificial Intelligence and Deepfakes in Strategic Deception Campaigns: The U.S. and Russian Experiences. In: Pashentsev, E. (eds) The Palgrave Handbook of Malicious Use of AI and Psychological Security. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-22552-9_19
[5] https://www.researchgate.net/publication/371463213_Artificial_Intelligence_and_Deepfakes_in_Strategic_Deception_Campaigns_The_US_and_Russian_Experiences
[6] https://yann.lecun.com/exdb/publis/pdf/lecun-iscas-10.pdf
[7] https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained
[8] Florian Schroff; Dmitry Kalenichenko; James Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering" (PDF). The Computer Vision Foundation. Retrieved 4 October 2023
[9] Taigman, Yaniv; Yang, Ming; Ranzato, Marc'Aurelio; Wolf, Lior (June 24, 2014), "DeepFace: Closing the Gap to Human-Level Performance in Face Verification", Conference on Computer Vision and Pattern Recognition (CVPR), Facebook Research Group
[10] https://hightech.fm/2017/05/31/ntechlab-findface-best
[11] roboticsandautomationnews.com/2023/07/06/vanceai-art-generator-guide-create-visuals-in-few-steps/69873/
[12] https://dl.acm.org/doi/abs/10.1145/3422622
[13] https://truthandtrustonline.com/wp-content/uploads/2020/10/TTO05.pdf
[14] https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html
[15] https://arxiv.org/abs/1901.08971
[16] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3564638
[17] https://revistainnovacion.com/nota/12321/es_la_biometria_el_futuro_de_la_autenticacion/
[18] https://therecord.media/cencora-data-breach-notification
[19] https://github.com/ondyari/FaceForensics
[20] https://github.com/ahaliassos/LipForensics
[21] https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981