Sistema para la detección de deepfake de sonido

Cristhian Eduardo Castillo Meneses
Kevin Gianmarco Zarama Luna

Universidad Icesi
Facultad de Ingeniería
Ingeniería de Sistemas
Cali

2020

Sistema para la detección de deepfake de sonido

Cristhian Eduardo Castillo Meneses
Kevin Gianmarco Zarama Luna

Proyecto de Grado

Tutor:
Christian Urcuqui, MSC

Universidad Icesi
Facultad de Ingeniería
Ingeniería de Sistemas
Cali

2020

Tabla de Contenido

Resumen **5**

Abstract **6**

Lista de acrónimos **7**

Glosario de términos **8**

Motivación y antecedentes **11**
Contexto 11
Antecedentes del problema 11
Justificación 12

Descripción del problema **13**

Objetivos del proyecto **14**
Objetivo general 14
Objetivos específicos 14

Marco teórico **15**
Inteligencia artificial 15
Aprendizaje automático 15
Red neuronal 15
Aprendizaje profundo 16
Aprendizaje no supervisado 16
Red generativa antagónica (GAN) 17
Redes Neuronales Convolucionales(CNN) 17
Deepfake 17
Espectrograma 17

Estado del arte **19**
DeepFake Audio Detection 19
Adversarial Audio Synthesis (WaveGAN) 20
DeepSonar: Towards Effective and Robust Detection ofAI-Synthesized Fake Voices 20

Generalization Of Audio Deepfake Detection 21
Método de detección de Deepfake mediante técnicas de *Machine Learning* 21
Matriz de estado del arte 21

Metodología **23**
Desarrollo del sistema 23
Metodologías ágiles 23
Prototipado 24
Stack de tecnologías 27
Despliegues 28
CRISP-DM (Cross-Industry Standard Process for Data Mining) . 29
Fases del modelo *CRISP-DM* 30
Esquema de trabajo 33
Análisis de riesgos y limitaciones 34
Cronograma 35

Experimentos y resultados **36**
Tecnologías usadas 36
Experimentos 36
Resultados 38

Contribución y entregables **39**
Contribuciones 39
Aportes relacionados con el objetivo del proyecto. 39
Aportes relacionados con el desarrollo de capacidades del investigador. 39
Entregables 39

Conclusiones y trabajo futuro del proyecto **40**
Conclusiones 40
Trabajo futuro 40

Resumen

Con el rápido crecimiento y aceptación que han tenido todos los ámbitos de la inteligencia artificial se han hecho muchos modelos que son capaces de crear imágenes, videos o audios que parecen cada vez más reales, lo que puede causar que personas con malas intenciones puedan hacer falsificaciones de noticias para obtener algún beneficio para ellos, por lo cual es importante crear mecanismos que permitan detectar contenido que no es auténtico y que ha sido generado por algún método de inteligencia artificial (*deepfake*).

En este proyecto se propone un método para la detección de *deepfakes* de sonido, haciendo uso del espectrograma de los audios para generar una representación visual de estos y su posterior procesamiento, con la técnica de propuesta, haciendo uso de redes neuronales para definir si el audio es creado mediante técnicas de inteligencia artificial o es realmente creado por una persona.

El método que se ha propuesto en este proyecto dio como resultado una exactitud de 92.23 % con los datos de prueba, lo cual evidencia que tiene un buen puntaje para poder diferenciar si un audio es auténtico o no y de este modo evitar que las personas puedan caer en estafas.

Abstract

With the rapid growth and acceptance of artificial intelligence, there has been a growing number of models that can create images, videos or audios that seem real, which can cause that people with bad intentions create fake news to obtain some benefit for the people with bad intentions, so it is important to create mechanisms that detect synthetic content and created with some artificial intelligence method (deepfake).

This project proposes a method for the detection of sound deepfakes, making use of the spectrogram of the audios to generate a visual representation of these and their subsequent processing, with neural networks to define whether the audio It is created using artificial intelligence techniques or if it is authentic.

The method that has been proposed in this project achieved an accuracy of 85% with the test data, which shows that it has a good score to be able to differentiate if an audio is authentic or not and thus has the potential to help people fall into scams.

Lista de acrónimos

GAN generative adversarial network.
AI artificial intelligence.
ML machine learning.
DL deep learning.
DNN deep neural network.
RGAs generative adversarial network.
CNN convolutional neuronal network.
MOS mean opinion score.
CRISP-DM Cross-Industry Standard Process for Data Mining.

Glosario de términos

Deepfake = son medios manipulados digitalmente y haciendo uso de inteligencia artificial, en la que una persona en una imagen, vídeo o audio existente se reemplaza por otra persona y suele presentar situaciones que no ocurrieron.

Ingeniería social = es un conjunto de técnicas que usan los cibercriminales para engañar a los usuarios incautos para que les envíen datos confidenciales, infecten sus computadoras con malware o abran enlaces a sitios infectados[1].

Reddit = sitio web de marcadores sociales y agregador de noticias donde los usuarios pueden añadir texto, imágenes, vídeos o enlaces.

Espectrograma = resultado de calcular el espectro de una señal por ventanas de tiempo de esta. Resulta una gráfica tridimensional que representa la energía del contenido frecuencial de la señal según va variando a lo largo del tiempo.

Escala mel = es una escala musical perceptual de tonos juzgados como intervalos equiespaciados por parte de los observadores.

Espectrograma mel = espectrograma en donde las frecuencias del sonido se transforman a una escala mel[2].

Función sigmoide = forma de describir cómo se realiza la transición de niveles bajos hasta niveles altos de una progresión temporal.

$$y = \frac{1}{1 + e^{-x}} \tag{1}$$

Scrum = marco de trabajo para desarrollo ágil de software que se ha expandido a otras industrias. Es un proceso en el que se aplican de manera regular un conjunto de buenas prácticas para trabajar colaborativamente, en equipo y obtener el mejor resultado posible de proyectos.

Dueño del producto = se asegura de que el equipo Scrum trabaje de forma adecuada desde la perspectiva del negocio. El dueño del producto ayuda al usuario a escribir las historias de usuario, las prioriza, y las coloca en el registro del producto.

Índice de figuras

1. Diagrama con la arquitectura que se presenta en el modelo de detección DeepFake Audio Detection[18]. 19
2. Prototipo del home realizado en Figma. 25
3. Prototipo del demo realizado en Figma. 26
4. Prototipo del paper realizado en Figma. 27
5. Flujo del sistema y stack de tecnologías usadas en el proyecto. 28
6. Esquema de despliegues continuos del sistema. 29
7. Diagrama del ciclo de vida de la metodología *CRISP-DM*[23]. 30
8. Los espectrogramas son representaciones visuales de sonido. Si se observa de cerca, se logra notar que la imagen de la izquierda es más borrosa que la imagen de la derecha. ¡Eso es porque el audio con el que se generó el espectrograma de la derecha es falso! 32
9. Arquitectura del modelo propuesto en el proyecto. 32
10. Cronograma del anteproyecto 35
11. Arquitectura del modelo realizado para el experimento 2, una capa *convLSTM* añadida al inicio del modelo. 37

Índice de cuadros

1. Tabla de comparación de los métodos detección de *deepfake* audio. 22
2. Tabla de distribución de los conjuntos de datos del conjunto de datos de ASVSpoof 2019[24]. 31
3. Tabla de riesgos identificados para el proyecto 34
4. Tabla de comparación de los resultados del modelo base y de los experimentos. 37
5. Tabla de comparación de los métodos detección de *deepfake* audio. 38

Motivación y antecedentes

Contexto

La palabra *deepfake* es el nombre que se le da al contenido que es originado mediante de inteligencia artificial, con el cual se puede realizar contenido que puede parecer real, pero es totalmente creado por inteligencia artificial. Con la de aparición de tecnologías como las redes generativas antagónicas (GAN)[3] en el 2014 por parte de Ian GoodFellow el *deepfake* se empezó a extender ampliamente. Lo que ha provocado que los modelos para realizar *deepfake* tanto de audios, videos e imágenes cada vez sean más realistas al punto de que no se pueda distinguir contenido real de uno sintético. Esto ha traído consigo una gran cantidad de noticias falsas, que en ocasiones también está acompañada de videos, imágenes o audios, que también pueden ser falsos, por lo que cada vez se vuelve más difícil distinguir contenido real de algo creado con el propósito de manipular.

La investigación en el campo del *deepfake* de audio no tiene la misma importancia y relevancia en la actualidad como lo tiene el *deepfake* de video e imágenes que se encuentra en un gran auge, esto a razón de que es más fácil encontrar documentación y conjuntos de datos, y también porque se crea mucho más contenido de *deepfake* de video o de imagen. Sin embargo, el *deepfake* de audio es igual de importante e incluso más peligroso que las otras variantes de *deepfake*, por su potencial de poner en situaciones de decisiones rápidas a las personas, por ejemplo, la llamada de una inteligencia artificial haciéndose pasar por el dueño de la compañía pidiendo que se realice una serie de pagos con el propósito de concretar unas alianzas.

Entonces se puede decir que nos encontramos en un momento donde el *deepfake* se están extendiendo y avanzando rápidamente y donde el *deepfake* de audio no está recibiendo la misma atención que sus pares de video e imágenes, a pesar de ser potencialmente peligroso.

Antecedentes del problema

El termino *deepfake* se originó en el 2017 dentro de Reddit luego de que un usuario llamado *"deepfake"*posteara haber desarrollado un algoritmo que permitía transponer rostros de celebridades porno[4]. A partir de ese momento empezaron a surgir muchas publicaciones bajo el concepto de

deepfake con diferentes variantes como el *deepfake* de audio, imágenes y video.

Bajo el concepto de *deepfake*, *BuzzFeed* publica un video falso de Barack Obama advirtiendo que entrabamos a una era donde nuestros enemigos pueden hacer que parezca que cualquiera pueda decir cualquier cosa[5]. Este tipo de contenido generado sintéticamente por inteligencia artificial puede usarse para la manipulación o para la ingeniería social.

Llegado a este punto podemos decir que el *deepfake* amenaza con volver muy difusa la línea entre la verdad y el engaño lo que lleva a que las personas no puedan diferenciar lo real de lo falso y se creen ambientes llenos de desconfianza.

Justificación

Es importante tener la capacidad identificar cuando un sonido ha sido creado o manipulado por medio de inteligencia artificial para poder diferenciarlo de sonidos reales y así evitar situaciones de chantaje, intimidación y sabotaje[6].

Descripción del problema

Vivimos en un mundo donde la información abunda al punto de saturarnos. Un efecto de esto es que hoy en día es difícil distinguir información digital verdadera de la que ha sido manipulada con malas intenciones, como es el caso de las noticias falsas.

La alteración de la información ha visto un gran aumento con la llegada de modelos de inteligencia artificial que facilitan esta tarea. Dichos modelos que se dedican a la falsificación de información se conocen como modelos de *deepfake*. La aparición de estos modelos se han vuelto un problema debido al mal uso que se le puede dar como falsificación de pruebas en juicios, difamación, creación de noticias falsas, entre otras aplicaciones mal intencionadas.

Objetivos del proyecto

Los objetivos que se plantearon al inicio del proyecto y en los cuales se siguió en el proyecto son:

Objetivo general

Desarrollar un sistema para la detección de contenido auditivo alterado.

Objetivos específicos

- Recopilar un conjunto de datos de datos reales y alterados.
- Proponer un método para detección de *deepfake* auditivo.
- Evaluar el método propuesto con otros existentes.

Marco teórico

El marco teórico con la recopilación de antecedentes e investigaciones previas para el proyecto es el siguiente:

Inteligencia artificial

Es una ciencia de la computación que se centra en la creación de programas y mecanismos para que las máquinas puedan mostrar comportamientos considerados inteligentes, y se puede definir como un sistema automatizado que es capaz de analizar los datos de su entorno y tomar decisiones de forma automática, para así maximizar sus posibilidades de éxito en algún objetivo o tarea. Esto se aplica cuando una máquina es capaz de realizar funciones asociadas a los seres humanos, como razonar, aprender o resolver problemas[7].

Aprendizaje automático

El aprendizaje automático o *machine learning*, es un subcampo de las ciencias de la computación y una de las grandes ramas de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan o mejoren su rendimiento. A lo que se refiere aprender es: cuando un agente mejora su desempeño con la experiencia, es decir que la habilidad de desarrollar alguna tarea no estaba en su genotipo al comienzo[8], por lo que se busca realizar que las máquinas de maneras heurísticas les permitan convertir muestras de datos en programas de computadora para que generalicen comportamientos e inferencias de un amplio conjunto de datos. Para la detección de *deepfakes* es muy común usar métodos de aprendizaje automático, de manera que se usan muchos de los avances que se han realizado en el campo del aprendizaje automatizado, tanto como para la creación como para la detección de los *deepfakes*[9].

Red neuronal

La red neuronal o también conocido como sistema conexionista es una parte del aprendizaje profundo que esta vagamente inspirado en el comportamiento observado en su homologo biológico. Consiste en un conjunto

de unidades, llamada neuronas artificiales, conectadas entre sí para transmitirse señales la información de entrada atraviesa la red neuronal donde produce valores de salida[10].

Aprendizaje profundo

Aprendizaje profundo o *deep learning*, es un conjunto de algoritmos de aprendizaje automático que intenta modelar abstracciones de alto nivel en datos usando arquitecturas computacionales que admiten transformaciones no lineales múltiples e iterativas de datos expresados en forma de matriz[11].

El aprendizaje profundo es una parte de un conjunto más amplio de métodos de aprendizaje automático basados en realizar representaciones de datos que a su vez están usando redes neuronales artificiales compuestas por varias capas en niveles, de los cuales el primer nivel, aprende algo simple y luego envía la información al siguiente nivel, el cual toma toda esa información sencilla y la combina para componer una información un poco más compleja y así sucesivamente. Mediante esta técnica se puede realizar la identificación del habla, de imágenes, generar los algoritmos que ayudan en la detección de los *deepfakes*. Sistemas como Siri y Cortana son potenciados, en gran parte, por los sistemas de aprendizaje profundo.

Aprendizaje no supervisado

Muchos de los métodos de creación de *deepfakes* son realizados en el aprendizaje no supervisado, ya que este tipo de aprendizaje nos permite no tener un conocimiento a priori, por lo cual solo se conoce los datos de entrada, pero no los datos de salida[12]. Son métodos de aprendizaje automático en el cual el aprendizaje no supervisado, mediante un conjunto de datos de entrada es tratado para que por medio de un modelo de aprendizaje nos dé una inferencia de un conjunto de variables aleatorias a resultados conocidos o etiquetados con un modelo de densidad para el conjunto de datos según similitudes o patrones entre los datos, lo cual hace que clasifique a los datos en grupos atendiendo a las variables de los datos[13].

Red generativa antagónica (GAN)

Red generativa antagónica o *generative adversarial network* (GAN), dado un conjunto de entrenamiento, en esta técnica aprende a generar nuevos datos con las mismas estadísticas que el conjunto de entrenamiento. Las GAN crean nueva información a través de un discriminador que se actualiza dinámicamente, por lo cual, se puede decir que el generador no está entrenado para generar nuevos datos para minimizar las diferencias entre los reales y los generados, sino que está entrenado para tratar de engañar al discriminador, lo cual permite que el modelo aprenda de una manera dinámica y no supervisada. Por ejemplo, una GAN puede generar nuevas fotografías de personas inexistentes que tengan rasgos auténticos para los observadores humanos[14].

Redes Neuronales Convolucionales(CNN)

Una CNN (*convolutional neuural network*) es un tipo de red neuronal artificial con aprendizaje no supervisado que procesa sus capas imitando al cortex visual del ojo humano para identificar distintas características en las entradas que en definitiva hacen que pueda identificar objetos. Este tipo de red neural contiene varias capas especializadas y con una jerarquía[15], esto quiere decir que aprende de las capas que contiene y procesan sobre la entrada formando un mapa de características abstractas como la salida. Este tipo de red neuronal forma matrices con un grupo de información, llamado kernel, y esto es una capa; a medida que se realicen más capas, la red es capaz de reconocer formas complejas.

Deepfake

Es un acrónimo de las palabras *fake* que quiere decir falsificación y *deep* learning. El *deepfakes* consiste en una técnica que permite editar videos, imágenes o audios falsos de personas creadas por los algoritmos de deep learning a partir de imágenes, videos o audios de personas reales.

Espectrograma

El espectrograma es la representación visual resultado de calcular el espectro de frecuencias de tramas de una señal que varía con el tiempo.

El gráfico estándar que se utiliza para mostrar el espectrograma tiene en un eje simboliza el tiempo y en el otro eje la frecuencia y opcionalmente en una tercera dimensión indica la amplitud de una frecuencia particular en un momento particular que está explicado por la intensidad o el color de cada punto de la imagen. Esto puede usarse para generar el *deepfake* de audio, mediante métodos de generación de imágenes que luego se transforman a audio, el espectrograma también puede usarse con los métodos de detección de *deepfakes* de audio, encuentran una diferencia en el audio generado por un humano y en el audio generado por un algoritmo de inteligencia artificial[16].

Estado del arte

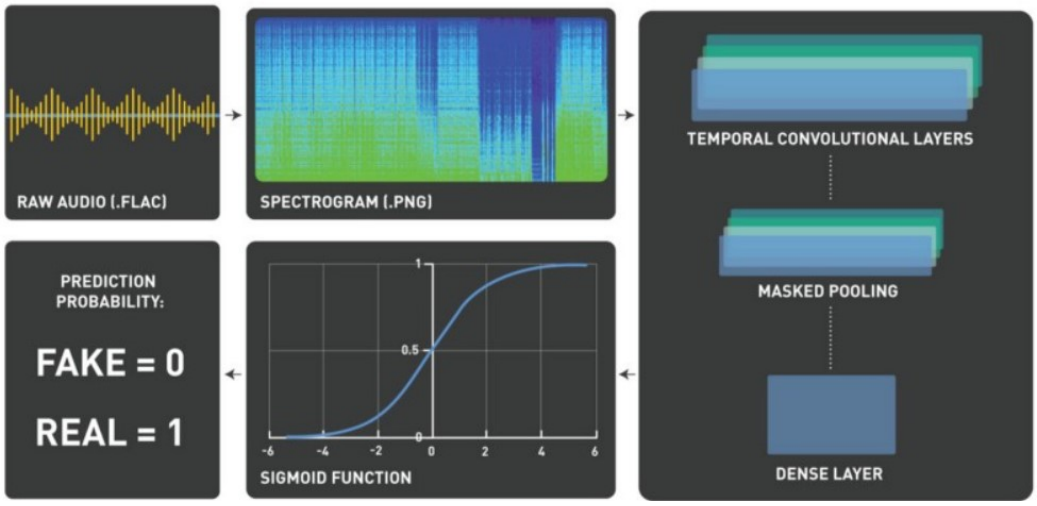
Para el estado del arte del proyecto, se describen 4 trabajos previos que presentan diferentes soluciones. Las cuales abordan el problema de diferentes maneras.

DeepFake Audio Detection

En este proyecto se presenta el modelo de detección de audio *deepfakes*, el cual presenta la solución para un desafío y hace uso de un conjunto de datos ASVSpoof de Google liberado en 2019 para fomentar el desarrollo de detección de audio *deepfake*, el desafío consiste en realizar modelos para la verificación automática de hablante[17], es decir la detección de audio *deepfake*.

El modelo del detector *deepfake* es una red neuronal profunda que utiliza la convolución temporal. Aquí hay una descripción general de alto nivel de la arquitectura del modelo:

Figura 1: Diagrama con la arquitectura que se presenta en el modelo de detección DeepFake Audio Detection[18].



Primero, el audio sin procesar se pre-procesa y se convierte en un espectrograma de frecuencia mel; esta es la entrada para el modelo. El mode-

lo realiza convoluciones sobre la dimensión de tiempo del espectrograma, luego usa agrupación enmascarada para evitar el sobreajuste. Finalmente, la salida se pasa a una capa densa y una función de activación sigmoidea, que finalmente genera una probabilidad entre 0 (falso) y 1 (real)[18].

Los resultados que obtuvo este modelo tuvieron una exactitud de 99% en entrenamiento, 95% en validación y un 85% en prueba, se puede evidenciar que los resultados de este modelo fueron muy altos, por lo que se puede decir que es un método de detección muy fiable y seguro.

Adversarial Audio Synthesis (WaveGAN)

En este trabajo nos presentan el modelo de *WaveGAN* y nos expone el uso de GANs para la generación y detección de contenido de *deepfake* de audio. Expone la posibilidad de ver las señales de audio como series de tiempo. En cada momento en el tiempo el sonido puede representarse como un valor, por lo que se puede ver el sonido (o amplitud a) como una función de tiempo t. Debido a esta posibilidad de ver las señales de audio como una serie de tiempo, podemos representar un audio como un vector.

Por lo que podemos decir que $a = f(t)$. Donde la función f es continua, por naturaleza, por lo que, para convertir una señal en un conjunto finito de números, se debe elegir una frecuencia de muestreo.

También plantea que el uso de convoluciones transpuestas para escalar el audio desde el ruido a un vector tendrá un impacto en el audio sintetizado, ya que notoriamente produce artefactos de tablero de ajedrez en las imágenes y su equivalente en señales de audio. Por lo que el discriminador podría aprender fácilmente a detectar audio falso basándose solo en este artefacto, deteriorando todo el proceso de entrenamiento. Para esto en su modelo de *WaveGAN* propone una solución llamada mezclado de fase, evitando convoluciones transpuestas y usando capas de muestreo superior (vecinos más cercanos) seguidas de convoluciones normales[19].

DeepSonar: Towards Effective and Robust Detection ofAI-Synthesized Fake Voices

En este artículo construyen una red neuronal profunda para discernir las voces falsas sintetizadas por inteligencia artificial. Hacen uso de la función de activación de neuronas por capas con la conjetura de que pueden

capturar las diferencias sutiles entre las voces falsas reales y sintetizadas por IA, al proporcionar una señal más limpia a los clasificadores que las entradas sin procesar. Sus experimentos se llevan a cabo sobre 3 conjuntos de datos (FoR, Sprocket-VC y MC-TT) que contienen datos en inglés y chino[20].

Generalization Of Audio Deepfake Detection

En este artículo hacen uso de una red neuronal residual con función de perdida LMCL y aumento de enmascaramiento de la frecuencia en línea para forzar a la red neuronal a aprender más sobre la incorporación de características robustas. El modelo se evaluó sobre los datos de acceso lógico de ASVspoof 2019 y también sobre su versión que contiene ruido con el propósito de simular escenarios más realistas. En el modelo que plantean reemplazan *softmax* con LCML, lo que les da unos mejores resultados. Por lo que plantean que LCML es capaz de forzar al modelo para aprender características más robustas que tienen una mejor capacidad de generalización[21].

Método de detección de Deepfake mediante técnicas de *Machine Learning*

En este artículo de proyecto de grado del semestre 2020-1 de los estudiantes de la Universidad Icesi se presentan varios métodos de detección de *deepfakes* de video, los cuales ya han sido expuestos por otras personas y también se llevó a cabo un sistema de detección de *deepfake* propio por parte de los estudiantes el cual está presentado a detalle en el artículo que ellos mismos hicieron y comparaciones entre los métodos de detección de *deepfakes* de video que ellos encontraron y el que desarrollaron[22].

Matriz de estado del arte

En la *tabla 1* se realiza un resumen de lo que se encuentra en el estado del arte.

Dentro de los criterios de comparación se encuentran:

- **Tecnología usada:** este criterio dice que tipo de tecnología fue usada en el artículo para poder identificar audio alterado.

- **Acceso a código fuente:** este criterio dice si el código fuente se encuentra disponible para el público en general.
- **Acceso a los datos:** este criterio indica si los datos que usaron para generar los experimentos se encuentran disponibles para el público general.
- **Uso de IA:** el uso de inteligencia artificial en general ha mostrado buenos resultados para este tipo de problemas, por lo que este criterio indica el uso de inteligencia artificial para hacer la detección.
- **Métricas:** indica las métricas usadas para la evaluación de los mecanismos.

Tabla 1: Tabla de comparación de los métodos detección de *deepfake* audio.

Método / Propiedad	**DeepFake Audio Detection**	**WaveGAN**	**DeepSonar**	**Generalization Of Audio Deepfake Detection**	**DeepAudio Detector**
Tecnología Usada	Redes neuronales convolucionales	Redes generativas antagónicas	Redes neuronales profundas	Redes neuronales residuales	Redes neuronales convolucionales
Acceso a código fuente	Si	Si	No	No	Si
Métricas	Exactitud 85 %	Puntaje SC09 4.7	Exactitud 98.1 %	Puntaje EER 1.26 %	Exactitud 92.23 %

Metodología

Para la metodología del proyecto se decidió hacer uso de *CRISP-DM* y metodologías agiles. Esta decisión se tomó debido a que se necesita una parte de ciencia de datos por lo cual se escogió a *CRISP-DM*, el cual es una metodología estándar para el proceso de ciencia de datos y para la parte de gestión del proyecto se ha elegido hacer uso de las metodologías agiles.

Desarrollo del sistema

Metodologías ágiles

Las metodologías agiles permiten adaptar la forma de trabajo a las condiciones del proyecto, debido a que se pueden dar varias iteraciones del ciclo de vida del proyecto y su esquema de trabajo es incremental, consiguiendo flexibilidad y la facilidad de respuesta para adaptar el proyecto a los cambios que puedan suceder, debido a que se desarrolló esta metodología con el principio de que los requisitos y soluciones pueden evolucionar con el tiempo según la necesidad del proyecto.

Para el proyecto se decidió realizar una iteración la cual está basada en la metodología ágil por cada una de las entregas:

- **Planificación** En esta etapa se realizó la planificación de los objetivos del proyecto en la iteración que se llevó a cabo en su desarrollo, teniendo en cuenta los entregables que se debían realizar, para cumplir con cada uno de los cursos de proyecto de grado.
- **Diseño** Para la etapa de diseño se realizó un diseño preliminar de la forma de trabajar que se iba a seguir a lo largo del proyecto y como iban a ser los entregables y acordar las reuniones de equipo y el trabajo que debía realizar cada uno de los integrantes y también se realizó un diseño y la propuesta del diseño que se debía seguir para cumplir con los objetivos.
- **Codificación** Se realizó la codificación del modelo que se había propuesto en la fase anterior y se realizó la fase de entrenamiento y se obtuvieron los resultados.

- **Pruebas** Se realizó una inspección de los resultados que se obtuvieron en el modelo y se sacaron las debidas conclusiones con respecto a los resultados y los objetivos del proyecto.
- **Entregas** Para realizar las entregas basadas en los hitos y compromisos con nuestro tutor. Estas entregas e hitos estaban alineadas con cumplir los objetivos del proyecto.
- **Documentación** En esta fase se realizó la documentación final del proyecto, en la cual se describió el proyecto en este documento.

Prototipado

Se construyó un prototipo en figma para que sirviera como guía para el desarrollo del sistema.

Figura 2: Prototipo del home realizado en Figma.

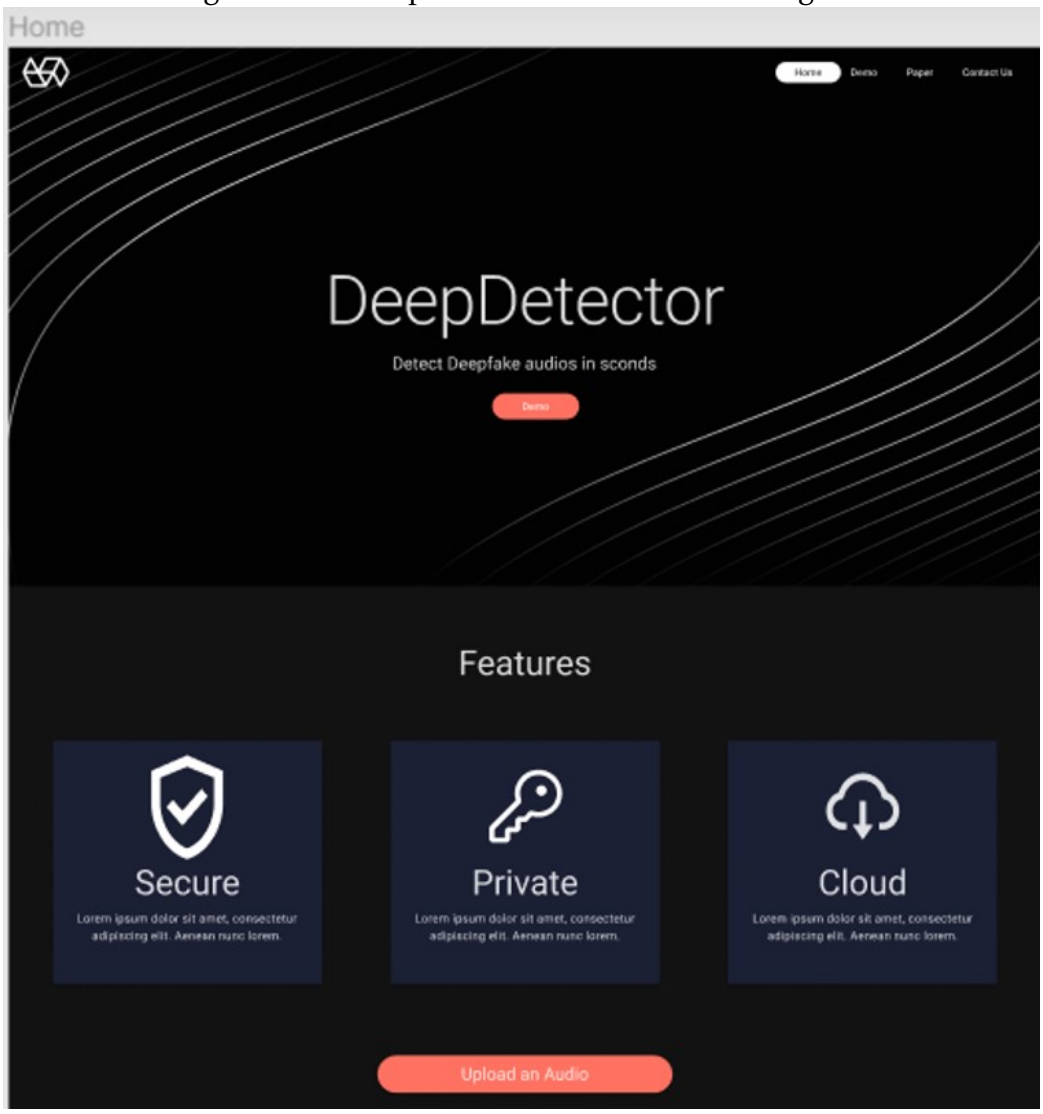


Figura 3: Prototipo del demo realizado en Figma.

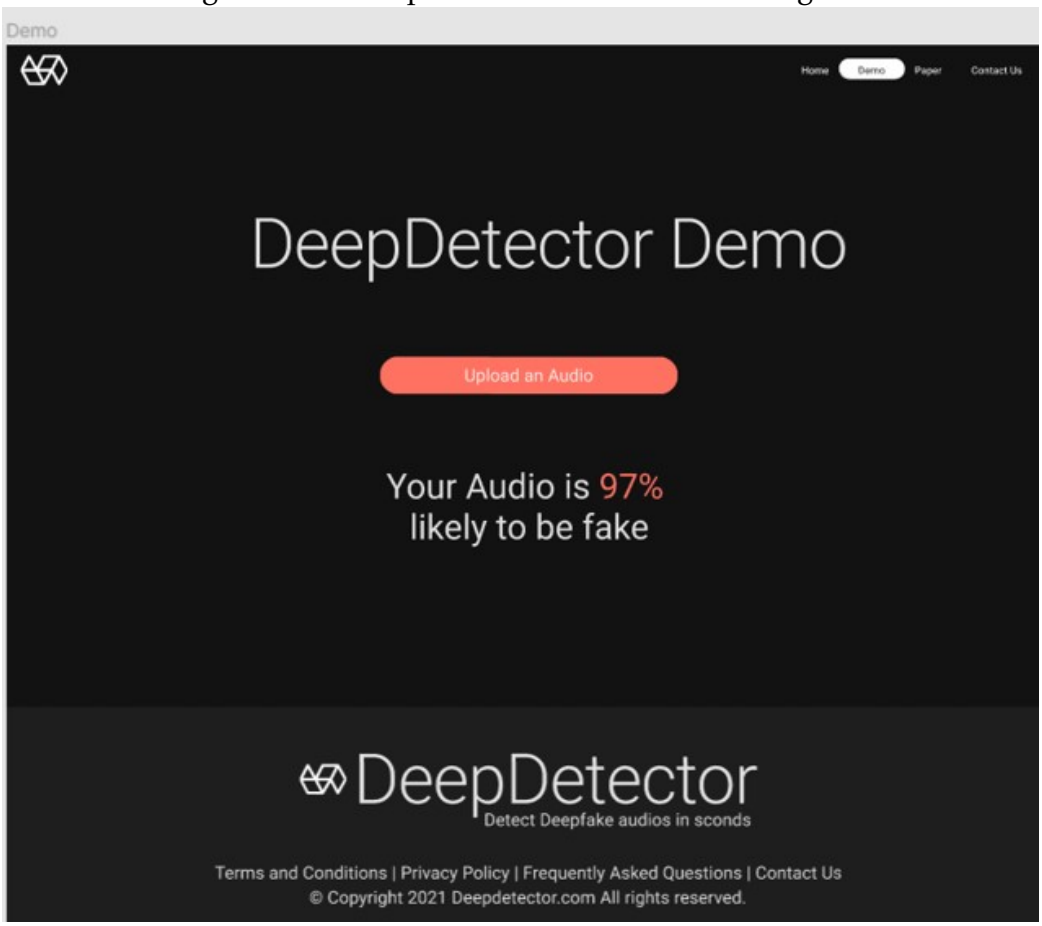
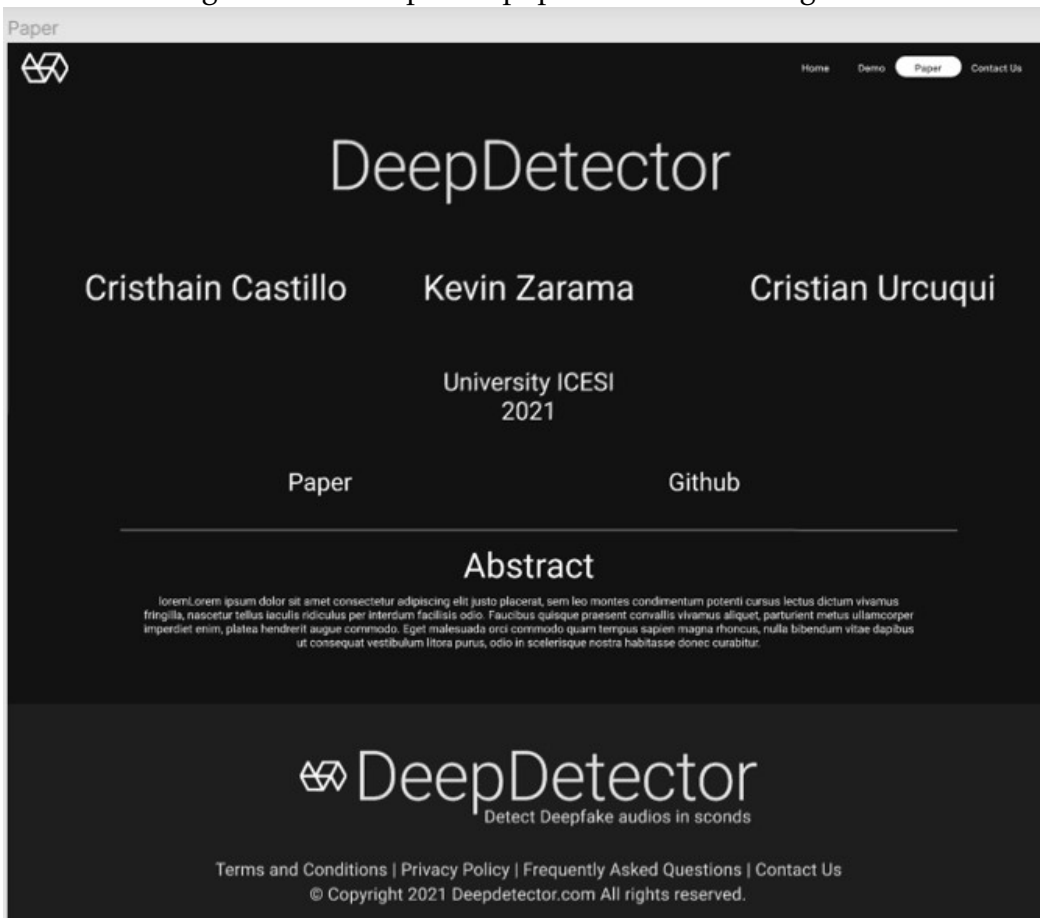


Figura 4: Prototipo del paper realizado en Figma.



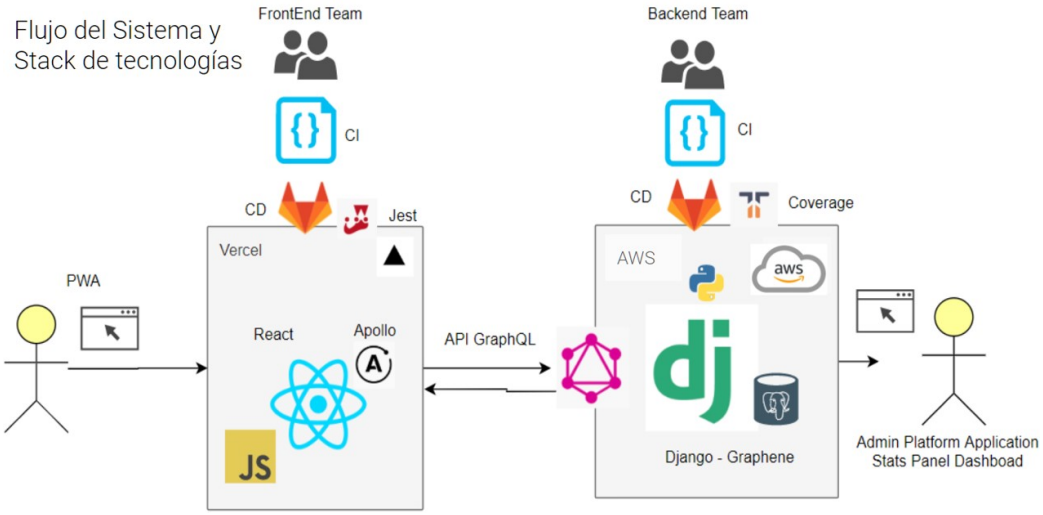
Stack de tecnologías

Se decidió desarrollar apoyándose del sistema de Gitlab CD/CD.

En el Frontend, se decidió utilizar React Js junto a Apollo Client para conectarse a la API de GraphQL se se construyó el backend. Para los despliegues se utilizó el servicio de Vercel.

Por otro lado, en el backend se decidió utilizar Python con Django para construir un API GraphQL y utilizar los servicios de AWS para su despliegue.

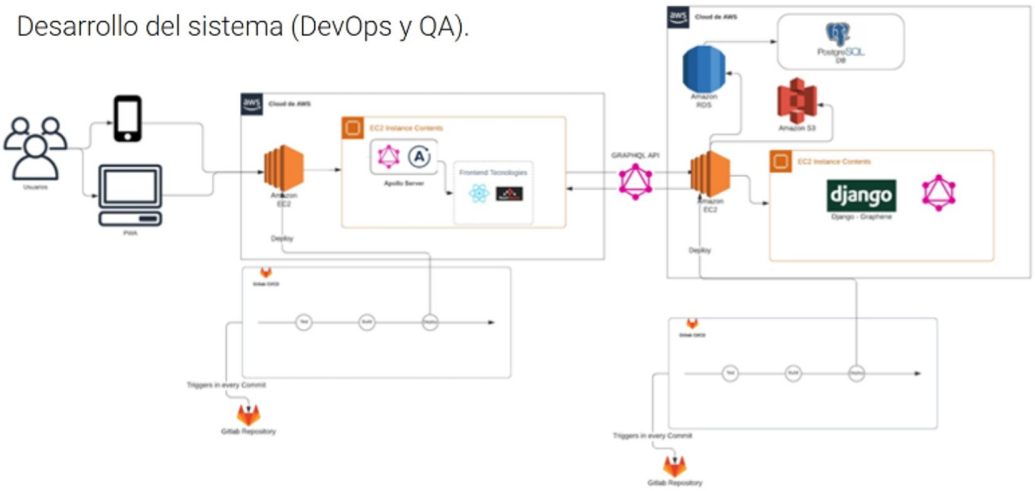
Figura 5: Flujo del sistema y stack de tecnologías usadas en el proyecto.



Despliegues

Haciendo uso de Gitlab se implementaron prácticas de DevOps con el propósito de despliegues e integraciones continuas que redujeran nuestros tiempos de desarrollo. Por lado de Frontend los despliegues eran a Vercel y de Backend a un entorno en AWS compuesto por un sistema de S3 para almacenamiento, un EC2 para el API y un RDS para la base de datos.

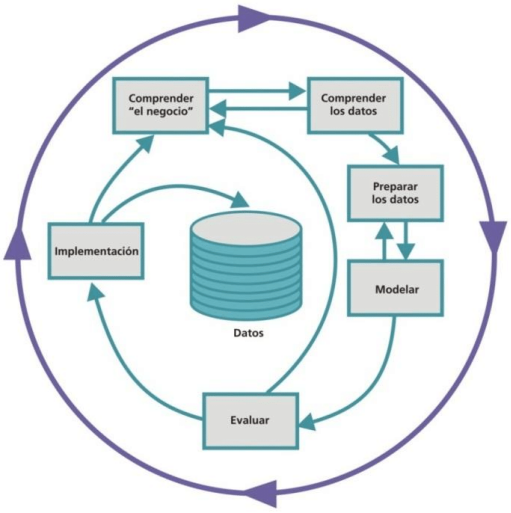
Figura 6: Esquema de despliegues continuos del sistema.



CRISP-DM (Cross-Industry Standard Process for Data Mining)

Es una metodología ampliamente usada en los proyectos que tienen que ver con el análisis y procesamiento de grandes cantidades de datos o *big-data*. El modelo CRISP-DM contempla todas las fases de un proyecto, que no son necesariamente rígidas, también sus respectivas tareas y las relaciones entre estas tareas[23].

Figura 7: Diagrama del ciclo de vida de la metodología *CRISP-DM*[23].



Fases del modelo *CRISP-DM*

- **Comprensión del negocio** Para esta fase se decidió que se determinaran los objetivos de la minería de datos y un diseño preliminar para lograr los objetivos. Se estudiaron diferentes métodos que detectan los *deepfakes* de sonido, e identificaron las estrategias que usaba cada uno de los métodos de detección. Se eligió una de las estrategias que se presentan el uno de los modelos llamado *fake voice detection*. Para discernir entre audio real y falso, el detector usa representaciones visuales de clips de audio llamados espectrograma, el cual además se usa para entrenar el modelo de síntesis de voz. Para los datos se seleccionó el conjunto de datos de Google´s 2019 ASVSpoof dataset[24], liberado en el 2019 por parte de Google en el desafío ASVSpoof para el desarrollo de modelos de detección de *deepfake* de sonido.

- **Comprensión de los datos** En esta fase se comenzó la recolección de los datos necesarios para lograr el objetivo del proyecto, para continuar con su comprensión y la identificación de los problemas que puedan tener. Luego, ya seleccionado el modelo con el cual se iba a realizar el proyecto, se procedió a la comprensión de los datos del conjunto de datos de 2019 ASVSpoof, con lo cual se observó que el

conjunto de datos contiene más de 133.000 clips de audio entre audios reales y falsos, con oradores hombres y mujeres.

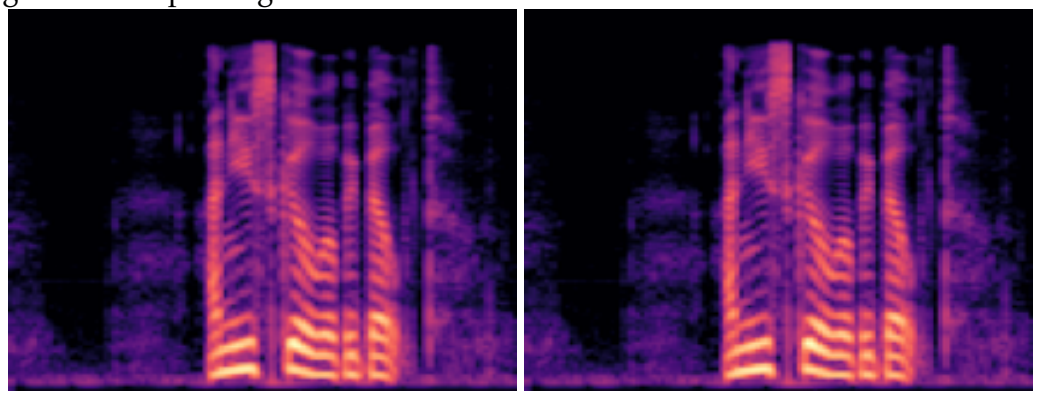
Tabla 2: Tabla de distribución de los conjuntos de datos del conjunto de datos de ASVSpoof 2019[24].

Subconjunto	# Oradores		# Declaraciones			
	Hombre	Mujer	Acceso lógico		Acceso físico	
			Auténticos	Falsos	Auténticos	Falsos
Entrenamiento	8	12	2580	22800	5400	48600
Desarrollo	8	12	2548	22296	5400	24300

El conjunto de datos ASVspoof 2019 para acceso lógico se basa en una base de datos estándar de síntesis de voz de varios hablantes llamada VCTK2[25]. Las declaraciones genuinas se recopilan de 107 hablantes (46 hombres, 61 mujeres) y sin efectos significativos de ruido de canal o de fondo. Las declaraciones falsificadas se generan a partir de datos genuinos utilizando varios algoritmos de suplantación de identidad diferentes. El conjunto de datos completo se divide en tres subconjuntos, el primero para entrenamiento, el segundo para desarrollo y el tercero para evaluación. El número de hablantes en los dos subconjuntos anteriores se ilustra en la tabla 2.

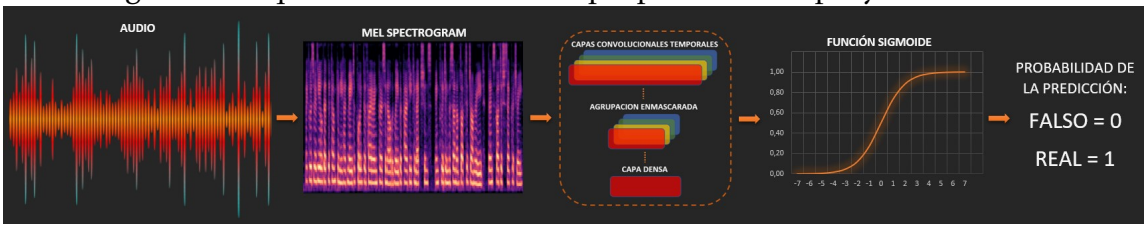
- **Análisis de los datos** En esta fase se seleccionaron los datos del conjunto de datos de Google´s 2019 ASVSpoof, no se procedió a realizar limpieza de los datos, debido a que estos ya estaban preparados específicamente para este fin en el desafío, por lo cual se usó la totalidad de los datos del conjunto de datos (dimensiones especificadas anteriormente). Finalmente se procedió a realizar el espectrograma mel de cada uno de los audios del conjunto de datos para poder introducirlos en el modelo.

Figura 8: Los espectrogramas son representaciones visuales de sonido. Si se observa de cerca, se logra notar que la imagen de la izquierda es más borrosa que la imagen de la derecha. ¡Eso es porque el audio con el que se generó el espectrograma de la derecha es falso!



- **Modelado** Para esta fase se realizó un modelo en el cual, se basa en usar la arquitectura propuesta por parte de Dessa IA lab, publicado el 28 de septiembre de 2019, en su proyecto *deepfake audio detection*, en el cual se hace uso de redes neuronales profundas compuesta por capas de convolución temporal, la cual se describe en la sección del estado del arte, la entrada del modelo son los espectrogramas mel; primero el modelo toma el espectrograma y realiza convoluciones que crea matrices bidimensionales, las cuales son muy efectivas para tareas de visión artificial, como en la clasificación y la segmentación de imágenes; para después usa una agrupación enmascarada para evitar el sobreajuste del modelo, finalmente, la salida se pasa a una capa densa y posteriormente a una función de activación sigmoide, para así finalmente obtener una probabilidad de predicción entre 0 y 1, en donde 0 significa un audio falso y 1 un audio real.

Figura 9: Arquitectura del modelo propuesto en el proyecto.



En esta fase se realizaron dos experimentos, los cuales son: el primer experimento consistió en realizar un conjunto de datos más grande para el entrenamiento y las pruebas del modelo, mezclando el conjunto de datos con que originalmente se hizo el entrenamiento y pruebas del modelo, el conjunto de datos de ASVSpoof2019, con una muestra del conjunto de datos de *The fake-or-real dataset* de APTLY lab; y el segundo experimento consistió en adicionar capas de ConvLSTM a la arquitectura del modelo que se tenía en un principio para observar añadir una dimensión temporal en el modelo ayudaba a mejorar los resultados del modelo; los experimentos que se realizaron se describen con detalle en los experimentos que se presenta en la sección de resultados y experimentos.

- **Evaluación** En la fase de evaluación se obtuvieron los resultados con el criterio de exactitud. El porcentaje que se obtuvo como resultado de correr el modelo con la arquitectura que se realizó para el proyecto y corriendo un total de 5 épocas con el conjunto de datos de ASVSpoof2019, fue de 98.5% en entrenamiento, 95.2% de validación y 82.3% en prueba.

- **Implementación** En la fase de implementación, se llevó a cabo el ordenamiento de los datos que se obtuvieron a partir del modelo y se realizó una revisión final para ver si se cumplían todos los objetivos que se habían propuesto a inicios del proyecto.

  Para la implementación del proyecto se hizo uso de tecnologías tales como GraphQL para poder manejar las peticiones y las consultas, Django para manejar el backend, React para el frontend, Apollo para manejar las peticiones del frontend hacia el API, entre otros.

Esquema de trabajo

Para el desarrollo de actividades y el seguimiento por parte del tutor, mientras se esté en el curso de proyecto de grado 1, se han realizado reuniones cada 15 días para tratar los temas necesarios y discutir sobre los entregables y después mientras se esté en proyecto de grado 2 se decidió hacer las reuniones una vez en la semana para concretar y dejar más actividades para lograr el objetivo del proyecto. Los incrementos y las validaciones del producto serán realizadas por el tutor del proyecto.

Análisis de riesgos y limitaciones

Se han identificado los siguientes riesgos que podrían afectar la ejecución del proyecto. A continuación, se presentan dichos riesgos y su descripción.

Tabla 3: Tabla de riesgos identificados para el proyecto

Riesgo	**Efecto**	**Mitigación**
Falta de conocimientos para el despliegue de modelos de AI	Retraso en el cronograma y retraso en la fase de análisis de información	Adelantar paralelamente el desarrollo de software y el aprendizaje de los modelos
Ausencia de alguno de los integrantes por fuerza mayor	Problemas para el otro integrante y perdida de información del integrante ausente	Comunicación efectiva y conocimiento de las actividades y el progreso de cada uno
Hardware disponible por los integrantes del grupo no suficiente para el entrenamiento de los modelos	Retraso en el cronograma	Tratar de realizar con tiempo de anticipación el entrenamiento de los modelos
Métricas no satisfactorias en los modelos	Retraso en el cronograma	Elegir varios modelos que tengan un buen desempeño con el audio o modelos que ya hayan sido usados para el *deepfake* de sonido

Cronograma

Figura 10: Cronograma del anteproyecto

Tareas	Fecha de Inicio	Fecha de Fin
Proyecto de Grado	**jueves, 13 de agosto de 2020**	**viernes, 25 de junio de 2021**
Ante Proyecto	**jueves, 13 de agosto de 2020**	**viernes, 4 de diciembre de 2020**
Primera Entrega	**jueves, 13 de agosto de 2020**	**viernes, 25 de septiembre de 2020**
Definir Título del Proyecto	jueves, 13 de agosto de 2020	domingo, 23 de agosto de 2020
Escribir la motivación y antecedentes	domingo, 23 de agosto de 2020	miércoles, 2 de septiembre de 2020
Definir el problema	miércoles, 2 de septiembre de 2020	sábado, 12 de septiembre de 2020
Definir los objetivos	sábado, 12 de septiembre de 2020	martes, 22 de septiembre de 2020
Segunda Entrega	**viernes, 25 de septiembre de 2020**	**viernes, 30 de octubre de 2020**
Correcciones de la primera entrega	viernes, 25 de septiembre de 2020	lunes, 5 de octubre de 2020
Hacer Marco teórico	lunes, 5 de octubre de 2020	jueves, 15 de octubre de 2020
Hacer Estado del arte	jueves, 15 de octubre de 2020	domingo, 25 de octubre de 2020
Tercera Entrega	**viernes, 30 de octubre de 2020**	**viernes, 4 de diciembre de 2020**
Realizar analisis de riesgos	viernes, 30 de octubre de 2020	martes, 10 de noviembre de 2020
Definir Cronograma	martes, 10 de noviembre de 2020	sábado, 21 de noviembre de 2020
Definir la métología	sábado, 21 de noviembre de 2020	jueves, 3 de diciembre de 2020
Proyecto de Grado II y Producto	**viernes, 4 de diciembre de 2020**	**viernes, 25 de junio de 2021**
Prototipo	**viernes, 4 de diciembre de 2020**	**domingo, 3 de enero de 2021**
Definir Colores del Producto	viernes, 4 de diciembre de 2020	lunes, 14 de diciembre de 2020
Crear los componentes Reútilizables	lunes, 14 de diciembre de 2020	jueves, 24 de diciembre de 2020
Realizar protótipo	jueves, 24 de diciembre de 2020	domingo, 3 de enero de 2021
Incremento 1	**domingo, 3 de enero de 2021**	**miércoles, 17 de febrero de 2021**
Definir infraestructura y técnologías	domingo, 3 de enero de 2021	jueves, 14 de enero de 2021
Construir una infraestructura de pruebas (Devops)	jueves, 14 de enero de 2021	lunes, 25 de enero de 2021
Realizar los componentes principales (Frontend)	lunes, 25 de enero de 2021	viernes, 5 de febrero de 2021
Recopilar Datasets de datos reales y alterados	viernes, 5 de febrero de 2021	martes, 16 de febrero de 2021
Incremento 2	**miércoles, 17 de febrero de 2021**	**viernes, 19 de marzo de 2021**
Probar Modelos para detectar Deepfake Audio	miércoles, 17 de febrero de 2021	jueves, 4 de marzo de 2021
Construir frontend del producto	jueves, 4 de marzo de 2021	viernes, 19 de marzo de 2021
Incremento 3	**viernes, 19 de marzo de 2021**	**domingo, 18 de abril de 2021**
Definir Modelo y entrenarlo.	viernes, 19 de marzo de 2021	sábado, 3 de abril de 2021
Construir backend del producto	sábado, 3 de abril de 2021	domingo, 18 de abril de 2021
Incremento 4	**domingo, 18 de abril de 2021**	**martes, 18 de mayo de 2021**
Mejorar el rendimiento del modelo	domingo, 18 de abril de 2021	lunes, 3 de mayo de 2021
Comparar modelo construido con los otros modelos probados	lunes, 3 de mayo de 2021	martes, 18 de mayo de 2021
Incremento 5	**martes, 18 de mayo de 2021**	**jueves, 17 de junio de 2021**
Escribir Resultados	martes, 18 de mayo de 2021	miércoles, 2 de junio de 2021
Escribir conclusiones	miércoles, 2 de junio de 2021	jueves, 17 de junio de 2021

Experimentos y resultados

Tecnologías usadas

Para el desarrollo del proyecto se usó Python3, junto con Jupyter notebook y las librerías de numpy, sklearn, scipy, utils_model, PIL, keras, tensorflow, tqdm, matplotlib, functools, skopt, pydub, entre otras.

Experimentos

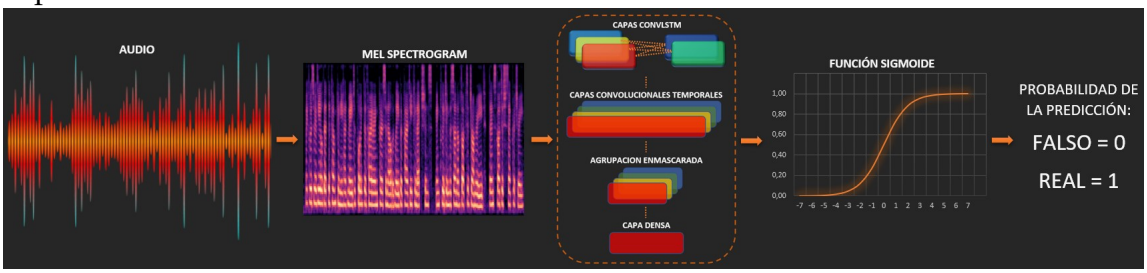
Experimento 1: En este experimento se mezcló el conjunto de datos con el cual se realizó el proceso de entrenamiento y pruebas para el modelo, asvspoof 2019, con una muestra del conjunto de datos de *The fake-or-real dataset (FoR)* de APTLY lab, que tiene más de 195.000 declaraciones de personas reales y de audios generados por computadores[26]. La muestra que se tomó consistió en 8000 audios reales y 8000 audios falsos.

La mezcla de los conjuntos de datos fue con la finalidad de aumentar los datos con los cuales estaba trabajando el modelo, también para aumentar el número de datos reales para que pasara del 10 % a un 25 % de datos reales en los datos de entrenamiento y también para observar cómo se comportaba el modelo con nuevos datos. Con esta mezcla de los conjuntos de datos se obtuvo una Exactitud del 99.25 % en entrenamiento, 96.78 % en validación y 87.8 % en pruebas.

Experimento 2: Para este experimento se adicionaron capas de ConvLSTM a la arquitectura del modelo, esto con la hipótesis de que en la dimensión temporal de los espectrogramas es una característica muy relevante para la clasificación de las imágenes como reales o falsas se agrega una Capa de ConvLSTM al inicio del modelo.

Las capas ConvLSTM son capas que se utilizan frecuentemente para predicciones espaciotemporal por su estructura convolucional en las transiciones de entrada a estado y de estado a estado. Estas características se usan para procesar series temporales de imágenes.

Figura 11: Arquitectura del modelo realizado para el experimento 2, una capa *convLSTM* añadida al inicio del modelo.



Con este experimento se consigue una Exactitud 99.8 % en entrenamiento, 97.25 % en validación y 92.23 % en el subconjunto de pruebas.

A continuación, se muestra un resumen de los resultados y características evaluadas en cada experimento.

Tabla 4: Tabla de comparación de los resultados del modelo base y de los experimentos.

Experimento	Descripción	ACC Train	ACC Val	ACC Test
Modelo base	Modelo con la arquitectura base	98.51 %	95.2 %	82.32 %
Experimento 1	Modelo con el aumento del conjunto de datos	99.25 %	96.78 %	87.8 %
Experimento 2	Adición de las capas LSTM convolucionales	99.8 %	97.25 %	92.23 %

En la tabla 4 anterior se presentan los resultados que se obtuvieron de hacer el entrenamiento, la validación y las pruebas del modelo que se tenía en un principio, y los dos experimentos que se realizaron, de esto se puede observar que mezclando el conjunto de datos original de ASVSpoof2019 con el conjunto de datos de *the real-or-fake* aumento el puntaje resultado del modelo que se tenía en un principio, y que el resultado mejoró aún más al adicionar las capas de convLSTM a la arquitectura, con lo cual se tiene que obtener una dimensión temporal si ayudó a mejorar los resultados del modelo y este fue el mejor resultado que se obtuvo.

Resultados

Los resultados que se obtuvieron luego de realizar el entrenamiento y los experimentos del modelo comparados con los resultados que han sido obtenidos por otros modelos fueron los siguientes:

Tabla 5: Tabla de comparación de los métodos detección de *deepfake* audio.

Método / Características	**DeepFake Audio Detection**	**WaveGAN**	**DeepSonar**	**Generalization Of Audio Deepfake Detection**	**DeepAudio Detector**
Tecnología Usada	Deep Neural Network	GANs	DNN	Residual neural network	CNN
Acceso a código fuente	Si	Si	No	No	No
Métricas	Exactitud 85 %	Puntaje SC09 4.7	Exactitud 98.1 %	Puntaje EER 1.26 %	Exactitud 92.23 %

Es de vital importancia contar con un conjunto de datos balanceado o usar técnicas para contrarrestar estos problemas al afrontar una tarea de clasificación de datos reales de falsos. El realizar el aumento del conjunto de datos para contrarrestar el problema de balanceo, presenta un aumento importante en la Exactitud con respecto al entrenamiento con el conjunto de datos original.

Hacer uso de capas LSTM convolucionales en problemas de clasificación audios con una representación en imágenes presenta resultados igualmente buenos como lo hace en la clasificación de otros tipos de imágenes.

El mejor resultado que se obtuvo con el modelo fue en el experimento 2, lo cual se puede evidenciar en la tabla 4, en el experimento 2 se adicionó capas de convLSTM a la arquitectura, obteniendo un 92.23 % de exactitud en el conjunto de datos de prueba, el cual está comparado en la tabla 5 con los otros resultados de los otros métodos de detección de *deepfake* que se investigaron a lo largo del proyecto.

Contribución y entregables

Contribuciones

Aportes relacionados con el objetivo del proyecto.

La idea de realizar este proyecto se origina en un proyecto que se desarrolló con anterioridad, el cual consiste en la detección de *deepfake*[22], en este proyecto se quiso dar un paso más con la detección de audios falsos y así poder tener más seguridad al decidir si un video es falso o verdadero. Se contribuye con una herramienta que permite distinguir entre contenido de audios generados con *deepfake* y el contenido real. La herramienta dicha anteriormente consta de un API GraphQL que recibe el audio y devuelve como resultado un número entre 0 (Falso) y 1 (Real). También con un método de detección mejorado frente a otros en la clasificación de audios *deepfake* y reales

Aportes relacionados con el desarrollo de capacidades del investigador.

Este proyecto de grado nos permitió como investigadores e Ingenieros de *Software* utilizar todos los procesos relacionados con la ciencia de datos en la metodología de *CRISPDM* y del ciclo del desarrollo de *Software* que aprendimos durante nuestra formación profesional, la cual dejamos constancia en el desarrollo de este proyecto.

Entregables

Como entregables del proyecto se tiene:

- El conjunto de datos usados en el proyecto.
- Documentación del proyecto.

Conclusiones y trabajo futuro del proyecto

Conclusiones

En este proyecto se propuso una herramienta que permite clasificar audios reales o falsos a partir de una red neuronal convolucional, durante la elaboración de esta herramienta se llega a las conclusiones:

- El uso de espectrogramas mel es una buena aproximación para resolver el problema de clasificar audios reales y falsos, ya que en el peor de los experimentos con el modelo base sin capas LSTM y un conjunto de datos desbalanceado logró obtener una exactitud mayor al 80 %, aunque se encuentra lejos de las mejores propuestas que alcanzan una exactitud 98 %, es una solución viable si no se cuenta con grandes recursos para implementar soluciones mucho más robustas que requieren un poder computacional mucho mayor.
- Al enfrentarse a los problemas de clasificación de audios alterados es de vital importancia contar con un conjunto de datos amplio y que esté lo más balanceado posible.
- La dimensión temporal en los espectrogramas para la clasificación de audios es una característica que cuenta con un gran peso.
- La etapa de preprocesamiento de los datos, es decir de la generación de espectrogramas es la etapa más difícil y que más tiempo lleva debido a que los espectrogramas deben estar construidos de tal forma que puedan pasar por las diferentes capas de un modelo convolucional. Esto es un problema que no se ve en otros tipos de clasificaciones con imágenes puesto que los espectrogramas provienen de un audio que cuenta con una dimensión temporal.

Trabajo futuro

Durante el desarrollo de este proyecto de grado hubo hipótesis que quedaron planteadas y prácticas de desarrollo que por el alcance del proyecto se decidieron omitir, por eso, lo planteamos como trabajo a futuro:

- Verificar si el modelo planteado y los que existen en el estado del arte tienen los mismos resultados con audios en otros idiomas diferentes al inglés como el español o portugués.

- Encontrar los hiper parámetros óptimos que presenten los mejores resultados en el modelo.
- Integrar el API con el Frontend generando así una herramienta que pueda ser accesible para cualquier tipo de público.
- Usar prácticas de tareas asíncronas como colas de tareas para que los tiempos de respuesta del API no afecten la experiencia de usuario.

Referencias

[1] Kaspersky. *Ingenieria social: definición*. URL: https://latam.kaspersky.com/resource-center/definitions/what-is-social-engineering.

[2] Leland Roberts. *Understanding the mel spectrogram*. 2020. URL: https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville y Yoshua Bengio. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[4] BBC Bitesize. *Deepfakes: What Are They and Why Would I Make One?* 2019. URL: https://www.bbc.co.uk/bitesize/articles/zfkwcqt.

[5] Craig Silverman. *How To Spot A Deepfake Like The Barack Obama–Jordan Peele Video*. 2018. URL: https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed.

[6] Danielle K. Citron y Robert Chesney. «Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security National Security». En: (2019).

[7] Salesforce. *Inteligencia Artificial: ¿Qué es? - Blog de Salesforce*. 2017. URL: https://www.salesforce.com/mx/blog/2017/6/Que-es-la-inteligencia-artificial.html.

[8] Oracle. *Definición de aprendizaje automático*. URL: https://www.oracle.com/co/data-science/machine-learning/what-is-machine-learning/.

[9] Aws amazon. *Aprendizaje automático*. URL: https://aws.amazon.com/es/machine-learning/what-is-ai/.

[10] AMD. *Entender el aprendizaje automático y el aprendizaje profundo*. 2020. URL: https://www.amd.com/system/files/documents/Machine-Learning-Primer.pdf.

[11] Raúl Arrabales. *Deep Learning: qué es y por qué va a ser una tecnología clave en el futuro de la inteligencia artificial*. 2016. URL: https://www.xataka.com/robotica-e-ia/deep-learning-que-es-y-por-que-va-a-ser-una-tecnologia-clave-en-el-futuro-de-la-inteligencia-artificia.

[12] Telefónica Tech. *Tipos de aprendizaje en Machine Learning: supervisado y no supervisado*. 2017. URL: https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/.

[13] Aprende IA. *¿Cómo funcionan las Convolutional Neural Networks? Visión por Ordenador*. URL: https://aprendeia.com/aprendizaje-no-supervisado-machine-learning/.

[14] Jason Brownlee. *Una introducción suave a las redes adversarias generativas (GAN)*. 2019. URL: https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/.

[15] Aprende machine learning. *Aprendizaje no Supervisado*. URL: https://www.aprendemachinelearning.com/como-funcionan-las-convolutional-neural-networks-vision-por-ordenador/.

[16] Luis Colomer. *Capítulo 10. Análisis espectral de los sonidos musicales*. URL: http://cursodeacusticamusical.blogspot.com/2016/02/capitulo-10-analisis-espectral-de-los.html.

[17] Yamagishi et al. *ASVspoof*. Inf. téc. URL: https://www.asvspoof.org/.

[18] IA lab Dessa. *Detecting Audio Deepfakes With AI*. 2019. URL: https://medium.com/dessa-news/detecting-audio-deepfakes-f2edfd8e2b35.

[19] Miller P. Donahuc D. McAuley J. «Adversarial audio synthesis». En: 2019.

[20] Wan et al. «DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices». En: (2020).

[21] Chen et al. «Generalization of Audio Deepfake Detection». En: (2020).

[22] Santiago Gutierrez Bolaños Bayron Daymiro Campaz Hurtado Juan David Diaz Monsalve. «Método de detección de Deepfake mediante técnicas de Machine Learning». En: (2020).

[23] Julio Villena Román. *CRISP-DM: La metodología para poner orden en los proyectos*. 2016. URL: https://www.sngular.com/es/data-science-crisp-dm-metodologia/.

[24] Yamagishi et al. *ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database*. Inf. téc. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019.

[25] MacDonald K Veaux C. Yamagishi J. *SUPERSEDED - CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit*. Inf. téc. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2017.

[26] APTLY lab. *The Fake-or-Real dataset*. Inf. téc.