
UNIVERSIDAD ESAN

FACULTAD DE INGENIERÍA

INGENIERÍA DE TECNOLOGÍAS DE INFORMACIÓN Y SISTEMAS

# Detección de audio deepfake en Español utilizando redes neuronales profundas: Análisis de tono, timbre y patrones de voz

Trabajo de investigación para el curso de Trabajo de Tesis I

Antonio Jesús Barrera Benetres

Asesor: Marks Calderón

Lima, 18 de noviembre de 2024

# Resumen

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ac odio tempor orci dapibus ultrices in iaculis nunc sed. Vivamus arcu felis bibendum ut tristique et egestas quis ipsum. Odio morbi quis commodo odio aenean sed adipiscing diam donec. Donec ultrices tincidunt arcu non sodales neque sodales ut. Fusce ut placerat orci nulla pellentesque dignissim enim sit amet. Facilisi etiam dignissim diam quis enim lobortis. Sit amet justo donec enim diam vulputate ut pharetra. Gravida in fermentum et sollicitudin ac orci phasellus egestas. Ultricies tristique nulla aliquet enim tortor at auctor. Nullam vehicula ipsum a arcu cursus vitae congue mauris. Convallis posuere morbi leo urna molestie at elementum eu facilisis. Elit at imperdiet dui accumsan sit amet nulla. Amet consectetur adipiscing elit pellentesque habitant morbi tristique senectus et. Mauris in aliquam sem fringilla ut morbi. Ultricies integer quis auctor elit sed vulputate mi sit. Nulla pellentesque dignissim enim sit amet venenatis urna cursus eget. Ac feugiat sed lectus vestibulum mattis ullamcorper. Eu augue ut lectus arcu bibendum. Rhoncus dolor purus non enim praesent elementum.

Nulla facilisi cras fermentum odio eu feugiat pretium. Massa massa ultricies mi quis hendrerit. Id leo in vitae turpis massa sed elementum. Quis vel eros donec ac odio tempor orci. Netus et malesuada fames ac turpis egestas integer eget aliquet. Velit ut tortor pretium viverra suspendisse potenti. Ut enim blandit volutpat maecenas. Nibh tellus molestie nunc non blandit. Mus mauris vitae ultricies leo integer malesuada nunc vel. Vel elit scelerisque mauris pellentesque pulvinar pellentesque habitant. Neque viverra justo nec ultrices dui sapien eget. Vitae aliquet nec ullamcorper sit. Dui id ornare arcu odio ut sem nulla pharetra diam. Et magnis dis parturient montes. Varius morbi enim nunc faucibus.

**Palabras claves:** uno, dos, tres, cuatro

# Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ac odio tempor orci dapibus ultrices in iaculis nunc sed. Vivamus arcu felis bibendum ut tristique et egestas quis ipsum. Odio morbi quis commodo odio aenean sed adipiscing diam donec. Donec ultrices tincidunt arcu non sodales neque sodales ut. Fusce ut placerat orci nulla pellentesque dignissim enim sit amet. Facilisi etiam dignissim diam quis enim lobortis. Sit amet justo donec enim diam vulputate ut pharetra. Gravida in fermentum et sollicitudin ac orci phasellus egestas. Ultricies tristique nulla aliquet enim tortor at auctor. Nullam vehicula ipsum a arcu cursus vitae congue mauris. Convallis posuere morbi leo urna molestie at elementum eu facilisis. Elit at imperdiet dui accumsan sit amet nulla. Amet consectetur adipiscing elit pellentesque habitant morbi tristique senectus et. Mauris in aliquam sem fringilla ut morbi. Ultricies integer quis auctor elit sed vulputate mi sit. Nulla pellentesque dignissim enim sit amet venenatis urna cursus eget. Ac feugiat sed lectus vestibulum mattis ullamcorper. Eu augue ut lectus arcu bibendum. Rhoncus dolor purus non enim praesent elementum.

Nulla facilisi cras fermentum odio eu feugiat pretium. Massa massa ultricies mi quis hendrerit. Id leo in vitae turpis massa sed elementum. Quis vel eros donec ac odio tempor orci. Netus et malesuada fames ac turpis egestas integer eget aliquet. Velit ut tortor pretium viverra suspendisse potenti. Ut enim blandit volutpat maecenas. Nibh tellus molestie nunc non blandit. Mus mauris vitae ultricies leo integer malesuada nunc vel. Vel elit scelerisque mauris pellentesque pulvinar pellentesque habitant. Neque viverra justo nec ultrices dui sapien eget. Vitae aliquet nec ullamcorper sit. Dui id ornare arcu odio ut sem nulla pharetra diam. Et magnis dis parturient montes. Varius morbi enim nunc faucibus.

**Keywords:** uno, dos, tres, cuatro

Para mi X, Y,X

# Agradecimientos

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ac odio tempor orci dapibus ultrices in iaculis nunc sed. Vivamus arcu felis bibendum ut tristique et egestas quis ipsum. Odio morbi quis commodo odio aenean sed adipiscing diam donec. Donec ultrices tincidunt arcu non sodales neque sodales ut. Fusce ut placerat orci nulla pellentesque dignissim enim sit amet. Facilisi etiam dignissim diam quis enim lobortis. Sit amet justo donec enim diam vulputate ut pharetra. Gravida in fermentum et sollicitudin ac orci phasellus egestas. Ultricies tristique nulla aliquet enim tortor at auctor. Nullam vehicula ipsum a arcu cursus vitae congue mauris. Convallis posuere morbi leo urna molestie at elementum eu facilisis. Elit at imperdiet dui accumsan sit amet nulla. Amet consectetur adipiscing elit pellentesque habitant morbi tristique senectus et. Mauris in aliquam sem fringilla ut morbi. Ultricies integer quis auctor elit sed vulputate mi sit. Nulla pellentesque dignissim enim sit amet venenatis urna cursus eget. Ac feugiat sed lectus vestibulum mattis ullamcorper. Eu augue ut lectus arcu bibendum. Rhoncus dolor purus non enim praesent elementum.

Nulla facilisi cras fermentum odio eu feugiat pretium. Massa massa ultricies mi quis hendrerit. Id leo in vitae turpis massa sed elementum. Quis vel eros donec ac odio tempor orci. Netus et malesuada fames ac turpis egestas integer eget aliquet. Velit ut tortor pretium viverra suspendisse potenti. Ut enim blandit volutpat maecenas. Nibh tellus molestie nunc non blandit. Mus mauris vitae ultricies leo integer malesuada nunc vel. Vel elit scelerisque mauris pellentesque pulvinar pellentesque habitant. Neque viverra justo nec ultrices dui sapien eget. Vitae aliquet nec ullamcorper sit. Dui id ornare arcu odio ut sem nulla pharetra diam. Et magnis dis parturient montes. Varius morbi enim nunc faucibus.

# Índice general

# Índice de Figuras

# Índice de Tablas

# Capítulo 1

# PLANTEAMIENTO DEL PROBLEMA

## 1.1. Descripción de la Realidad Problemática

En la actualidad, el uso de tecnologías avanzadas, como la inteligencia artificial (IA), ha permitido la creación de contenido multimedia falsificado conocido como deepfake. Este contenido incluye videos, imágenes y, más recientemente, audios. Los deepfakes de audio han adquirido una relevancia particular debido a su capacidad para imitar voces humanas con una precisión asombrosa, lo que ha desencadenado una serie de problemas relacionados con la seguridad y la confianza en las comunicaciones digitales (**heidari2023**). Esta tecnología ha comenzado a ser utilizada para fraudes y suplantación de identidad, presentando un desafío significativo para la detección de audios manipulados, especialmente en español, donde las herramientas actuales de detección no están adecuadamente adaptadas para capturar las particularidades del idioma.

En Perú, el fraude mediante deepfakes de audio ha mostrado un aumento considerable en los últimos años. Un informe de la Policía Nacional del Perú (PNP) revela que más de un millón de soles han sido robados mediante la clonación de voces utilizando IA. En estos fraudes, los delincuentes suelen imitar la voz de familiares o amigos para solicitar dinero en situaciones de emergencia falsas, lo que resulta en un engaño muy difícil de detectar por las víctimas (**rojas2023**). Este incremento de fraudes plantea un problema general relacionado con la incapacidad de las herramientas actuales para detectar audios falsificados, lo que facilita la comisión de delitos en el ámbito digital.

Uno de los problemas específicos más críticos es la falta de un dataset en español que contemple variaciones regionales y voces manipuladas para entrenar modelos de redes neuronales profundas. La mayoría de los avances en la detección de deepfakes de audio se han

centrado en el inglés, lo que genera una brecha importante en la capacidad de los modelos para adaptarse a las características del habla en español, que presenta variaciones significativas en términos de patrones de voz, frecuencia fundamental y ritmo del habla (**heidari2023**). Esta falta de datos específicos en español dificulta la construcción de modelos robustos que puedan detectar deepfakes con precisión en este idioma.

Otro problema específico está relacionado con las limitaciones técnicas de las herramientas actuales para analizar variables clave como el tono de voz, timbre de voz y formantes. Los audios falsificados mediante deepfakes logran imitar estas características acústicas de manera casi perfecta, lo que confunde tanto a los sistemas de detección como a las personas. En particular, los deepfakes en español plantean desafíos únicos debido a las diferencias fonéticas con otros idiomas, lo que complica aún más la detección efectiva de manipulaciones (**amesquita2023**).

Un tercer problema específico se refiere al uso de deepfakes de audio en fraudes por suplantación de identidad en Perú, donde los delincuentes utilizan esta tecnología para hacer pasar sus voces por las de familiares o colegas en situaciones de emergencia o negociación. Estos fraudes, que afectan tanto a individuos como a empresas, se ven agravados por la dificultad de las técnicas actuales para analizar características como la prosodia, articulación, transiciones entre fonemas y ruidos de fondo. Estos factores son críticos para la identificación precisa de audios falsificados, ya que los deepfakes no siempre logran replicar con fidelidad estos aspectos del habla humana, pero los sistemas de detección existentes no están optimizados para capturarlos (**chen2024**).

A nivel global, se ha documentado un incremento en el uso de deepfakes para cometer fraudes, con pérdidas millonarias en varios países. En Hong Kong, por ejemplo, se reportó un caso en el que un trabajador fue engañado mediante una videollamada con deepfakes de varios miembros de la junta directiva de su empresa, lo que resultó en un fraude de 25 millones de dólares (**chen2024**). Este tipo de casos destaca la urgencia de desarrollar soluciones tecnológicas más efectivas que permitan la detección de deepfakes de audio en español, donde las herramientas actuales siguen siendo insuficientes.

En el contexto peruano, la proliferación de fraudes basados en deepfakes de audio no solo afecta a los individuos, sino también a figuras públicas. Un caso reciente involucró a la presidenta Dina Boluarte, cuya voz fue manipulada mediante IA para hacer parecer que promovía una inversión fraudulenta, lo que generó confusión entre el público y demostró el poder de esta tecnología para influir en la opinión pública y causar daño reputacional (**amesquita2023**).

Por todo lo anterior, es evidente que la detección de deepfakes de audio en español requiere un enfoque más sofisticado. El desarrollo de modelos basados en redes neuronales pro-

fundas que puedan analizar variables acústicas clave, como el tono de voz, timbre de voz, patrones de voz, frecuencia fundamental, duración y ritmo del habla, formantes, nivel de energía del habla, ruidos de fondo, prosodia, articulación y transiciones entre fonemas, es esencial para mitigar el impacto de los fraudes por deepfakes en Perú. Estas variables juegan un papel fundamental en la autenticidad del habla y pueden proporcionar pistas valiosas para identificar audios manipulados. Sin embargo, la falta de modelos especializados en español y la escasez de datasets específicos siguen siendo los principales obstáculos para lograr una detección precisa y confiable.

## 1.2. Formulación del Problema

### 1.2.1. Problema General

El incremento del fraude en Perú mediante el uso de tecnologías deepfake de audio ha evidenciado la falta de herramientas adecuadas para detectar estos fraudes, especialmente en español. Las técnicas actuales no logran identificar eficazmente las características acústicas del español, como el tono de voz, timbre de voz, patrones de voz, frecuencia fundamental (pitch), duración y ritmo del habla, formantes, nivel de energía del habla (intensidad), ruidos de fondo, prosodia, articulación y transiciones entre fonemas, lo que facilita la suplantación de identidad y el fraude en las comunicaciones personales y empresariales.

### 1.2.2. Problemas Específicos

- La falta de un dataset en español que incluya variaciones regionales y voces manipuladas dificulta el entrenamiento de modelos de redes neuronales profundas para detectar deepfakes de audio en español, debido a las diferencias en patrones de voz, frecuencia fundamental (pitch) y ritmo del habla.
- Las técnicas actuales no logran detectar las variaciones en el tono de voz, timbre de voz y formantes en audios en español, lo que disminuye la precisión en la identificación de audios manipulados.
- Los fraudes por suplantación de identidad mediante deepfakes de audio en Perú son difíciles de detectar con las técnicas actuales debido a la falta de análisis de prosodia, articulación, transiciones entre fonemas y ruidos de fondo, lo que incrementa el riesgo de fraude.

## 1.3. Objetivos de la Investigación

Para la formulación de los objetivos de la presente investigación se elaboró un «árbol de objetivos» (véase Anexo 2)

### 1.3.1. Objetivo General

Desarrollar un modelo basado en redes neuronales profundas que permita detectar deepfakes de audio en español mediante el análisis de variables clave como tono de voz, timbre de voz, patrones de voz, frecuencia fundamental, duración y ritmo del habla, formantes, nivel de energía del habla, ruidos de fondo, prosodia, articulación y transiciones entre fonemas, mejorando la precisión en la identificación de audios manipulados para mitigar fraudes por suplantación de identidad en Perú.

### 1.3.2. Objetivos Específicos

- Desarrollar un dataset específico en español, con variaciones regionales y voces manipuladas, para entrenar un modelo de redes neuronales profundas que detecte deepfakes de audio
- Implementar un modelo de redes neuronales profundas que analice el tono de voz, timbre de voz y formantes para mejorar la precisión en la detección de deepfakes de audio en español.
- Evaluar la eficacia del modelo de redes neuronales profundas en la detección de deepfakes de audio en contextos de fraude por suplantación de identidad en Perú, considerando prosodia, articulación, transiciones entre fonemas y ruidos de fondo.

## 1.4. Justificación de la Investigación

### 1.4.1. Teórica

Desde una perspectiva teórica, esta investigación aporta al campo de la inteligencia artificial y el procesamiento de señales de audio, profundizando en el análisis y la detección de deepfakes mediante el uso de redes neuronales profundas. Actualmente, la mayoría de los avances en detección de deepfakes de audio se han centrado en el inglés, lo que deja una

brecha significativa en el desarrollo de modelos efectivos para el idioma español (**heidari2023**). Esta investigación se propone cubrir esa brecha, explorando cómo las características fonéticas del español (como la frecuencia fundamental, prosodia, formantes, y ruidos de fondo) pueden ser integradas en modelos de detección de audio. El desarrollo teórico de esta investigación se basa en los principios del aprendizaje profundo (deep learning) y su aplicación al análisis acústico, ampliando las bases teóricas sobre cómo los modelos de redes neuronales pueden ser adaptados para trabajar con diferentes idiomas y contextos culturales. Además, al enfocarse en las características específicas del español, esta investigación contribuye al desarrollo de datasets y herramientas especializadas que pueden ser utilizados en investigaciones futuras, tanto en el ámbito académico como en la industria

### 1.4.2. Práctica

La investigación sobre la detección de audios deepfake en español utilizando redes neuronales profundas es esencial en el contexto actual de seguridad digital, especialmente en Perú, donde se ha registrado un incremento preocupante de fraudes utilizando esta tecnología. Las suplantaciones de identidad mediante la clonación de voces, como lo demuestran los más de 94 casos de estafa reportados en 2023, con pérdidas superiores a un millón de soles (**rojas2023**), subrayan la vulnerabilidad de los sistemas actuales de detección. A nivel práctico, esta investigación tiene un impacto directo en la mitigación de estos fraudes. Al desarrollar modelos adaptados a las características específicas del español, como el tono de voz, timbre de voz y patrones de voz, se busca ofrecer soluciones tecnológicas robustas que permitan la identificación temprana de audios manipulados, reduciendo así los riesgos financieros y personales para los usuarios. Además, instituciones gubernamentales y empresas peruanas podrán aplicar los resultados de esta investigación para mejorar sus sistemas de seguridad y protección contra la suplantación de identidad, brindando un entorno más seguro para las comunicaciones digitales

### 1.4.3. Metodológica

La elección de redes neuronales profundas como metodología principal para la detección de audios deepfake se fundamenta en su capacidad superior para procesar y analizar grandes volúmenes de datos complejos, como las características acústicas de la voz humana. Las redes neuronales profundas permiten extraer automáticamente patrones y características de los datos de audio, como el tono de voz, timbre de voz, patrones de voz, frecuencia fundamental, prosodia, entre otros, lo que las convierte en una herramienta eficaz para abordar la naturaleza compleja de los deepfakes (**heidari2023**). Además, estas redes tienen la capacidad de aprender

de manera no supervisada y de adaptarse a las particularidades del idioma español, lo que es crucial dado que la mayoría de los avances en detección de deepfakes se han desarrollado en inglés.

Metodológicamente, el enfoque basado en redes neuronales profundas es ideal porque permite trabajar con un conjunto de variables independientes que incluyen tanto aspectos del habla (como la articulación y las transiciones entre fonemas) como la calidad del entorno de grabación (como los ruidos de fondo). La capacidad de estas redes para analizar simultáneamente múltiples dimensiones del audio y encontrar patrones que escapan a la detección humana o métodos tradicionales es fundamental para enfrentar la creciente sofisticación de los fraudes con deepfakes de audio (**ji2024**).

Otra razón metodológica clave es que el uso de redes neuronales profundas facilita la construcción de modelos predictivos que mejoran su precisión con el tiempo a medida que se entrenan con más datos. Esta flexibilidad es esencial, ya que el desarrollo de datasets específicos en español, con variaciones regionales y diferentes características de la voz, permitirá entrenar a los modelos de manera más eficaz para identificar manipulaciones en contextos reales (**rojas2023**). Por lo tanto, la metodología propuesta no solo es adecuada para la detección precisa de deepfakes de audio, sino que también es escalable y adaptable a futuros avances en la tecnología de falsificación de audios.

## 1.5. Delimitación del Estudio

### 1.5.1. Espacial

La presente investigación se desarrollará principalmente en el contexto de Perú, donde se ha identificado un incremento significativo en los fraudes utilizando tecnologías de deepfake de audio. El enfoque estará en analizar audios en español y las variaciones regionales del idioma dentro del país, ya que las características fonéticas del español peruano presentan particularidades que deben ser consideradas al desarrollar modelos de detección de deepfakes. Además, la aplicación de los resultados será relevante para instituciones gubernamentales, empresas y usuarios individuales en Perú, que enfrentan amenazas crecientes en el ámbito de la seguridad digital.

### 1.5.2. Temporal

El estudio abarcará el periodo 2023-2024, un lapso en el cual se han observado incrementos importantes en el uso fraudulento de deepfakes de audio tanto a nivel global como en Perú. La recolección de datos, construcción del dataset y desarrollo del modelo de redes neuronales profundas se llevarán a cabo en este periodo. Asimismo, se buscará analizar los fraudes recientes, reportados en los últimos dos años, para entender los patrones y tendencias de los ataques con deepfakes en audio

### 1.5.3. Conceptual

La investigación se centra en la detección de deepfakes de audio mediante el uso de redes neuronales profundas. Los conceptos clave abordados incluyen las características acústicas del habla, como el tono de voz, timbre de voz, frecuencia fundamental (pitch), duración y ritmo del habla, prosodia, formantes, articulación y ruidos de fondo, entre otros. Además, se explorará cómo estas variables son manipuladas por tecnologías de deepfake para suplantar identidades. El estudio también se delimita a la detección en español, debido a las diferencias fonéticas y lingüísticas entre este idioma y otros en los que se ha enfocado la investigación previa, como el inglés.

## 1.6. Hipótesis

### 1.6.1. Hipótesis General

El uso de un modelo basado en redes neuronales profundas que analice las variables acústicas clave como tono de voz, timbre de voz, patrones de voz, frecuencia fundamental, duración y ritmo del habla, formantes, nivel de energía del habla, ruidos de fondo, prosodia, articulación y transiciones entre fonemas mejora significativamente la precisión en la detección de deepfakes de audio en español, reduciendo el riesgo de fraudes por suplantación de identidad en Perú.

### 1.6.2. Hipótesis Específicas

- La creación de un dataset en español que incluya variaciones regionales y voces manipuladas mejorará significativamente la capacidad de las redes neuronales profundas para

detectar deepfakes de audio en este idioma

- El análisis del tono de voz, timbre de voz y formantes mediante redes neuronales profundas aumentará la precisión en la detección de deepfakes de audio en español.
- El modelo de redes neuronales profundas será más efectivo en la detección de deepfakes en contextos de fraude en Perú al incluir el análisis de prosodia, articulación, transiciones entre fonemas y ruidos de fondo, en comparación con las técnicas actuales.

### 1.6.3. Matriz de Consistencia

A continuación se presenta la matriz de consistencia elaborada para la presente investigación (véase Anexo A.1).

# Capítulo 2

# MARCO TEÓRICO

## 2.1. Antecedentes de la investigación

En esta sección se presentarán diversos artículos de investigación que abordarán diversas técnicas y enfoques que se emplearon para afrontar problemas similares al de esta tesis. Asimismo, a continuación se presenta un cuadro resumen (véase Anexo B.1) de lo que se presenta en esta sección.

### 2.1.1. Acoustic Features Analysis for Explainable Machine Learning-Based Audio Spoofing Detection (pr˙dehghani2018copper)

#### 2.1.1.1. Planteamiento del Problema y objetivo

La creciente sofisticación de tecnologías de generación y manipulación de voz sintética plantea serios problemas para la autenticidad en sistemas de verificación de identidad por voz. Este problema es particularmente relevante en la prevención de fraudes y en la autenticidad de registros de voz, ya que las técnicas de falsificación de audio (deepfakes) han alcanzado niveles de realismo que engañan tanto a los seres humanos como a los sistemas automáticos. El objetivo de este estudio es desarrollar un sistema de detección de deepfake de audio mediante un enfoque de aprendizaje automático (ML) que no dependa de representaciones espectrales comunes, sino de un conjunto diverso de características acústicas diseñadas a mano. Este método busca mejorar la transparencia y precisión del proceso de detección, proporcionando explicaciones claras de las decisiones del modelo.

#### 2.1.1.2. Técnicas empleadas por los autores

En lugar de emplear redes neuronales profundas y representaciones basadas en espectrogramas, el estudio utiliza características acústicas diseñadas manualmente, tales como características espectrales, temporales y de frecuencia (e.g., coeficientes MFCC, cromaticidad y energía). Además, se aplican técnicas de Inteligencia Artificial Explicable (XAI) para interpretar los resultados, específicamente la técnica SHapley Additive exPlanations (SHAP), que ayuda a clarificar cuáles características son cruciales en la detección de deepfakes de audio.

#### 2.1.1.3. Metodología empleada por los autores

Para evitar sesgos de reconocimiento biométrico, los autores implementaron un protocolo de prueba independiente del sujeto, asegurando que el modelo no se entrenara en los mismos individuos que se utilizarían para la prueba. La metodología incluyó experimentos extensivos de evaluación en tres conjuntos de datos y el uso de SHAP para interpretar las decisiones del modelo, mostrando qué características acústicas tenían mayor peso en la detección.

#### 2.1.1.4. Base de datos

Se utilizaron tres bases de datos de audio deepfake: ASVSpoof2019, FakeAVCelebV2 y un conjunto de datos In-The-Wild. ASVSpoof2019 incluye grabaciones de audio auténticas y falsificadas con ataques de texto-a-voz (TTS), conversión de voz (VC) y regrabación de audio. FakeAVCelebV2 es un conjunto multimodal de video y audio con variedad en género y etnicidad. In-The-Wild se compone de grabaciones con condiciones acústicas variadas para evaluar la generalización del modelo.

#### 2.1.1.5. Resultados obtenidos

El modelo alcanzó una precisión de 89 % en ASVSpoof2019, 94.5 % en FakeAVCelebV2 y 94.67 % en In-The-Wild, mostrando un rendimiento robusto y una alta capacidad de generalización comparado con métodos de última generación. Los resultados de SHAP confirmaron que las características acústicas eran esenciales para la detección, demostrando que un enfoque de características hechas a mano puede ser efectivo en comparación con métodos basados en aprendizaje profundo, además de proporcionar transparencia en el proceso de toma de decisiones.

## 2.1.2. Beyond the Illusion: Ensemble Learning for Effective Voice Deepfake Detection (pr˙dehghani2018copper)

### 2.1.2.1. Planteamiento del Problema y objetivo

La proliferación de deepfakes de voz ha creado una necesidad urgente de métodos de detección precisos, ya que estos falsos audios pueden ser utilizados en actividades delictivas como suplantación de identidad y fraudes financieros. Este estudio plantea como objetivo desarrollar un modelo de aprendizaje en ensamblaje (ensemble learning) que combina múltiples redes neuronales para mejorar la precisión y robustez en la detección de deepfakes de voz, superando los modelos individuales al integrar sus ventajas.

### 2.1.2.2. Técnicas empleadas por los autores

El enfoque de ensamblaje incluye cuatro arquitecturas de redes neuronales: red neuronal recurrente (RNN), red neuronal convolucional 1D (CNN 1D), memoria a largo plazo (LSTM) y memoria a largo plazo convolucional (ConvLSTM). Para la extracción de características, se emplean los coeficientes MFCC, cromaticidad y tasa de cruce por cero, que son combinados en vectores de características y normalizados para mejorar la consistencia y precisión.

### 2.1.2.3. Metodología empleada por los autores

El flujo de preprocesamiento incluye la extracción de características, estandarización dimensional y normalización de datos. Los modelos individuales son entrenados en el conjunto de datos, y luego sus resultados se fusionan mediante una técnica de soft voting en el modelo de ensamblaje para obtener la clasificación final.

### 2.1.2.4. Base de datos

La investigación se lleva a cabo en el conjunto de datos Fake-or-Real, que se divide en cuatro sub-conjuntos (’for-original’, ’for-norm’, ’for-2sec’, y ’for-rerec’) basados en la duración del audio y la tasa de bits. El modelo también fue probado en el conjunto de datos ASVSpoof en sus iteraciones 2019 y 2021.

#### 2.1.2.5. Resultados obtenidos

El modelo ECN-MF propuesto alcanzó una precisión de 99.5 % en el sub-conjunto original del dataset Fake-or-Real, 98 % en el sub-conjunto combinado y hasta un 99.6 % con redes CNN individuales en otros sub-conjuntos. Este enfoque de ensamblaje demostró que la combinación de múltiples modelos mejora significativamente la precisión y resistencia de la detección de deepfakes de voz, superando los resultados obtenidos con modelos individuales.

### 2.1.3. Comprehensive Multiparametric Analysis of Human Deepfake Speech Recognition (pr˙dehghani2018copper)

#### 2.1.3.1. Planteamiento del Problema y objetivo

El rápido avance en deepfakes de audio representa una amenaza creciente para la seguridad y autenticidad en comunicaciones críticas. Este estudio evalúa la capacidad humana para detectar deepfakes de audio en condiciones realistas, sin advertencia previa, abordando una brecha en la literatura donde los participantes generalmente están conscientes de que serán expuestos a deepfakes. El objetivo principal es entender cómo la información previa y la calidad del deepfake afectan la detección.

#### 2.1.3.2. Técnicas empleadas por los autores

Se implementa una métrica de calidad específica para deepfake de audio, que categoriza los audios en función de su fidelidad. Los experimentos se estructuran en dos fases: (1) sin aviso previo de exposición a deepfakes y (2) con variaciones en la calidad del audio.

#### 2.1.3.3. Metodología empleada por los autores

El experimento se dividió en dos partes: en la primera, los participantes jugaron el juego ”Dos Verdades, Una Mentiraüsando audios de países, de los cuales uno era un deepfake no anunciado. En la segunda, se evaluó la capacidad de los participantes para identificar deepfakes al variar la calidad del audio, utilizando una métrica diseñada para medir la fidelidad de estos audios.

#### 2.1.3.4. Base de datos

Para los experimentos, se generaron audios específicos en idiomas checo y eslovaco, asegurando la adaptación de los modelos de síntesis a estos idiomas para reflejar situaciones reales de engaño auditivo en entornos diversos.

#### 2.1.3.5. Resultados obtenidos

La detección sin aviso previo tuvo una efectividad baja, con una precisión que varió del 67 % al 94 % según la calidad del audio. La calidad del audio deepfake influenció significativamente la capacidad de detección, y los resultados sugieren que audios deepfake de alta calidad pueden llegar a ser indetectables para el oído humano, resaltando la necesidad de sistemas automatizados para apoyar la detección humana.

### 2.1.4. Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models (pr˙dehghani2018copper)

#### 2.1.4.1. Planteamiento del Problema y objetivo

El estudio explora cómo mejorar la detección de deepfakes en aplicaciones basadas en audio, como asistentes virtuales y sistemas de verificación de voz, mediante modelos de aprendizaje profundo que trabajen con transformaciones de espectrogramas. Su objetivo es mejorar la detección de deepfakes en aplicaciones de voz en tiempo real, donde la precisión y velocidad de detección son cruciales.

#### 2.1.4.2. Técnicas empleadas por los autores

El enfoque utiliza una combinación de modelos basados en espectrogramas, como CNN, RNN y C-RNN, así como embeddings de modelos preentrenados (Whisper, Speechbrain). Los espectrogramas se transforman mediante tres métodos diferentes: Transformada de Fourier de Tiempo Corto (STFT), Transformada-Q Constante (CQT) y Transformada Wavelet (WT), lo que permite capturar diferentes variaciones en el contenido frecuencial de los audios.

#### 2.1.4.3. Metodología empleada por los autores

Los audios se segmentaron en intervalos de dos segundos, generando espectrogramas de cada segmento. Los embeddings obtenidos de los modelos preentrenados se clasificaron con una MLP, y se integraron en un modelo de ensamblaje para lograr el mejor rendimiento.

#### 2.1.4.4. Base de datos

Se utilizó el conjunto de datos ASVSpoof2019 para evaluar la precisión de la detección, logrando un Equal Error Rate (EER) de 0.03, lo cual posiciona al modelo como altamente competitivo dentro de los sistemas top del desafío ASVspoof.

#### 2.1.4.5. Resultados obtenidos

El sistema logró una precisión competitiva, mostrando el potencial de los enfoques basados en espectrogramas para captar inconsistencias en audios falsos. Los resultados resaltan que el enfoque de ensamblaje con modelos profundos es efectivo en la detección de deepfakes de audio.

### 2.1.5. Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models (pr˙dehghani2018copper)

#### 2.1.5.1. Planteamiento del Problema y objetivo

Con el desarrollo de modelos generativos de alta calidad para la síntesis de sonidos ambientales, surge la necesidad de métodos de detección para audios ambientales falsificados (deepfake). Este estudio busca abordar la falta de investigación en esta área, proponiendo un sistema para distinguir entre sonidos ambientales grabados y generados.

#### 2.1.5.2. Técnicas empleadas por los autores

El modelo utiliza embeddings generados con el sistema CLAP (Contrastive Language-Audio Pretraining), entrenado específicamente para capturar similitudes semánticas y acústicas en audios. El sistema implementa una red MLP para clasificar los sonidos como reales o falsos.

#### 2.1.5.3. Metodología empleada por los autores

El audio se procesa en segmentos de cuatro segundos, generando embeddings de VGG, CLAP y PANN para identificar patrones específicos en audios ambientales falsificados.

#### 2.1.5.4. Base de datos

La base de datos proviene del desafío DCASE 2023, con más de 6 horas de audio real y 28 horas de audio generado mediante síntesis de sonido Foley.

#### 2.1.5.5. Resultados obtenidos

La arquitectura propuesta obtuvo una precisión del 98 %, demostrando que los embeddings específicos para audio ambiental mejoran la detección de deepfakes, y superan los modelos estándar en la clasificación de sonidos ambientales falsos.

### 2.1.6. Detection of Deepfake Media Using a Hybrid CNN-RNN Model and Particle Swarm Optimization (PSO) Algorithm (pr˙dehghani2018copper)

#### 2.1.6.1. Planteamiento del Problema y objetivo

Con la creciente sofisticación de los deepfakes, especialmente en contenido multimedia como video y audio, existe una necesidad crítica de métodos efectivos de detección para prevenir fraudes y preservar la integridad de la información. Este estudio propone una estrategia de detección que combina redes neuronales convolucionales (CNN) y redes neuronales recurrentes (RNN) junto con el algoritmo de optimización de enjambre de partículas (PSO), con el objetivo de mejorar la precisión de los modelos de detección de deepfakes (Al-Adwan et al., 2024).

#### 2.1.6.2. Técnicas empleadas por los autores

La técnica principal incluye una combinación híbrida de CNN y RNN para extraer características tanto espaciales como temporales, mejorada con PSO para optimizar los parámetros del modelo y mejorar la precisión en la clasificación de deepfakes.

#### 2.1.6.3. Metodología empleada por los autores

La metodología involucra el uso de CNN para capturar características espaciales en el contenido visual de los videos, mientras que la RNN analiza la secuencia temporal de los cuadros. PSO se emplea para ajustar los hiperparámetros de ambos modelos, mejorando así la capacidad del modelo para diferenciar entre videos genuinos y falsificados.

#### 2.1.6.4. Base de datos

Se usaron los conjuntos de datos Celeb-DF y Deepfake Detection Challenge (DFDC), ambos ampliamente utilizados en la investigación de detección de deepfakes y que contienen miles de clips de video manipulados y no manipulados.

#### 2.1.6.5. Resultados obtenidos

El modelo alcanzó una precisión promedio del 97.26 % en Celeb-DF y del 94.2 % en DFDC, superando a otros métodos de vanguardia en la precisión y robustez de detección (Al-Adwan et al., 2024).

### 2.1.7. Efficient Deepfake Audio Detection Using Spectro-Temporal Analysis and Deep Learning (pr˙dehghani2018copper)

#### 2.1.7.1. Planteamiento del Problema y objetivo

Debido al rápido desarrollo de tecnologías de deepfake en el ámbito de audio, existe un riesgo creciente para la seguridad digital y la autenticidad de la información auditiva. Este estudio busca mejorar la detección de deepfakes de audio utilizando un enfoque de análisis espectro-temporal y modelos de aprendizaje profundo (Sunkari & Srinagesh, 2024).

#### 2.1.7.2. Técnicas empleadas por los autores

La investigación emplea redes neuronales convolucionales (CNN) para la extracción de características espectrales y redes neuronales recurrentes (RNN) o de memoria a largo plazo (LSTM) para capturar las dinámicas temporales en las señales de audio.

#### 2.1.7.3. Metodología empleada por los autores

Se usó un preprocesamiento detallado que incluye la remoción de silencios y la normalización de audio para mejorar la calidad de entrada. La metodología de análisis espectro-temporal convierte las señales de audio en representaciones de espectrograma, permitiendo al modelo identificar patrones únicos en audios manipulados.

#### 2.1.7.4. Base de datos

El modelo fue entrenado y evaluado en el conjunto de datos ADD2022, el cual incluye una variedad de audios reales y manipulados en diversas condiciones y escenarios de grabación.

#### 2.1.7.5. Resultados obtenidos

El modelo mostró una alta capacidad de detección con métricas de precisión, F1 y EER mejoradas, destacándose como una herramienta efectiva en la detección de deepfakes de audio (Sunkari & Srinagesh, 2024).

### 2.1.8. Exploring Green AI for Audio Deepfake Detection (pr˙dehghani2018copper)

#### 2.1.8.1. Planteamiento del Problema y objetivo

Los sistemas de detección de deepfake en audio actuales, aunque efectivos, presentan un elevado costo ambiental debido al consumo de energía que requieren para entrenarse y operar. Este estudio introduce un enfoque de ”Green AI”que utiliza algoritmos de aprendizaje automático tradicionales y modelos de autoaprendizaje supervisado (SSL) para reducir la huella de carbono asociada a la detección de deepfakes (Saha et al., 2024).

#### 2.1.8.2. Técnicas empleadas por los autores

Se utilizaron modelos de autoaprendizaje supervisado preentrenados y métodos clásicos de clasificación como regresión logística y redes neuronales poco profundas, minimizando así la necesidad de recursos computacionales intensivos.

#### 2.1.8.3. Metodología empleada por los autores

El modelo extrae representaciones de audio a partir de un modelo SSL preentrenado (wav2vec 2.0) y las utiliza en algoritmos de clasificación clásicos sin afinación adicional, optimizando la eficiencia energética y reduciendo significativamente el número de parámetros del modelo.

#### 2.1.8.4. Base de datos

El conjunto de datos ASVspoof 2019 LA fue empleado para evaluar el rendimiento de detección y eficiencia energética de la metodología propuesta.

#### 2.1.8.5. Resultados obtenidos

El enfoque alcanzó un EER del 0.90 % utilizando menos de 1,000 parámetros entrenables, logrando resultados competitivos con un impacto ambiental reducido en comparación con enfoques tradicionales (Saha et al., 2024)

### 2.1.9. FakeSound: Deepfake General Audio Detection (pr˙dehghani2018copper)

#### 2.1.9.1. Planteamiento del Problema y objetivo

La detección de deepfakes en audio general, que incluye sonidos ambientales y otros tipos de audio no lingüístico, ha recibido poca atención. Este estudio propone una metodología para detectar audio general manipulado y localizar las regiones falsificadas dentro del mismo, lo cual es crucial para combatir el uso malicioso de esta tecnología en la falsificación de evidencia o noticias falsas (Xie et al., 2024).

#### 2.1.9.2. Técnicas empleadas por los autores

El estudio propone un modelo de detección que emplea un modelo de audio preentrenado como sistema de referencia, acompañado de redes profundas para clasificar y localizar las regiones falsificadas en el audio.

#### 2.1.9.3. Metodología empleada por los autores

Se usó una tubería de manipulación automatizada que genera segmentos falsos mediante enmascaramiento y regeneración, los cuales luego se concatenan con el audio original. Esta técnica de inpainting permite crear deepfakes realistas que combinan audio genuino y manipulado.

#### 2.1.9.4. Base de datos

El conjunto de datos FakeSound, que incluye audios manipulados de distintos tipos y duraciones, se usó para evaluar la eficacia del modelo, proporcionando un benchmark innovador para la detección de deepfakes de audio general.

#### 2.1.9.5. Resultados obtenidos

El modelo propuesto superó el desempeño de los sistemas de vanguardia en detección de deepfakes de audio general, mostrando un rendimiento superior incluso al de evaluadores humanos (Xie et al., 2024).

### 2.1.10. Targeted Augmented Data for Audio Deepfake Detection (pr˙dehghani2018cop

#### 2.1.10.1. Planteamiento del Problema y objetivo

Las detecciones de deepfake en audio tienden a sobreajustarse a manipulaciones vistas en el entrenamiento, limitando la generalización a deepfakes no observados. Este estudio presenta una técnica de aumento de datos específica que genera pseudo-deepfakes para mejorar la robustez de los detectores de audio deepfake (Astrid et al., 2024).

#### 2.1.10.2. Técnicas empleadas por los autores

La técnica principal se basa en un aumento de datos inspirado en ataques adversariales, donde los datos de audio reales se modifican para crear pseudo-falsificaciones que obliguen al modelo a mejorar su capacidad de generalización.

#### 2.1.10.3. Metodología empleada por los autores

Se generaron pseudo-deepfakes cerca del límite de decisión del modelo, optimizando el entrenamiento al incluir ejemplos ambiguos de audio que dificultan la clasificación. La estrategia se implementó en dos arquitecturas avanzadas de detección de audio deepfake, AASIST y RawNet2.

#### 2.1.10.4. Base de datos

La investigación utilizó el conjunto de datos ASVspoof 2019 para evaluar cómo el aumento de datos afecta la robustez del modelo frente a ataques no observados.

#### 2.1.10.5. Resultados obtenidos

La técnica de aumento mejoró significativamente la capacidad de ambos modelos para detectar deepfakes desconocidos, reduciendo el EER y aumentando la precisión en la mayoría de los tipos de deepfakes evaluados (Astrid et al., 2024).

## 2.2. Marco Teórico

### 2.2.1. Introducción al Deepfake de Audio

La tecnología deepfake ha revolucionado la creación y manipulación de contenido audiovisual mediante la aplicación de algoritmos de aprendizaje profundo. Un deepfake se define como una forma avanzada de manipulación digital que utiliza redes neuronales profundas, principalmente redes generativas adversarias (GAN), para crear contenido falso, como imágenes, videos o audios, que pueden engañar a los espectadores al simular la apariencia o voz de una persona real (Heidari, Jafari Navimipour, Dag, & Unal, 2023). En particular, los deepfakes de audio han cobrado relevancia debido a su capacidad para clonar voces, reproduciendo patrones vocales como el tono, el timbre y la prosodia. Estos avances han sido utilizados tanto para fines de entretenimiento como para actividades maliciosas, como el fraude y la suplantación de identidad, especialmente en el contexto de la voz (Rojas Berríos, 2023).

El uso de deepfakes de audio en esquemas de fraude ha aumentado en países como Perú, donde se han reportado numerosos casos en los que los delincuentes emplean inteligencia artificial para replicar la voz de familiares de las víctimas. Utilizando modelos de texto-a-voz y

conversión de voz, estos audios logran imitar de manera realista la voz de una persona, engañando a las víctimas y logrando que envíen dinero bajo falsas emergencias. Esta técnica permite a los estafadores clonar con precisión voces a partir de grabaciones previas, aprovechando así la confianza de las víctimas en la veracidad de la voz escuchada (Oorloff et al., 2024; Al-Adwan, Alazzam, Al-Anbaki, & Alduweib, 2024).

### 2.2.2. Teoría y Fundamentos en Análisis de Voz

El análisis de voz se enfoca en descomponer y analizar las señales de audio para obtener características distintivas de cada persona. Estos análisis incluyen parámetros acústicos como el tono (pitch), que es la frecuencia fundamental de la voz, y el timbre, que describe la çalidad"de la voz que hace que cada persona suene única. Estas características son esenciales para identificar posibles manipulaciones en deepfakes de audio, ya que los modelos de síntesis de voz pueden tener dificultades para replicar estos parámetros con precisión (Yi, Wang, Tao, Zhang, & Zhao, 2023).

El tono y el timbre son solo algunos de los componentes importantes en el análisis de voz para la detección de deepfakes. Otros parámetros clave incluyen el ritmo, la intensidad, y la prosodia, que se refiere a la entonación y las variaciones rítmicas en el habla. La prosodia, por ejemplo, puede ser compleja de replicar en modelos de síntesis de voz, ya que implica variaciones naturales que dependen del estado emocional y del contexto de la comunicación. Estas características acústicas y su análisis en el dominio del tiempo y la frecuencia proporcionan un marco para la detección de manipulaciones, permitiendo la identificación de patrones no naturales en audios generados artificialmente (Heidari et al., 2023; Lanzino et al., 2024).

### 2.2.3. Teoría y Fundamentos en Análisis de Voz

La detección de deepfakes ha avanzado significativamente con el desarrollo de redes neuronales profundas, que son capaces de aprender patrones complejos en datos no estructurados como el audio. Las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN) son modelos ampliamente utilizados para procesar datos secuenciales, como el audio, y extraer características relevantes para clasificar los audios en reales o falsos (Groh, Sankaranarayanan, Singh, Kim, Lippman, & Picard, 2024). Las CNN son efectivas en la extracción de patrones de frecuencia y textura, mientras que las RNN, especialmente los modelos de memoria a corto y largo plazo (LSTM), permiten capturar dependencias temporales en la señal de audio.

Recientemente, los investigadores han explorado modelos híbridos que combinan CNN y RNN junto con técnicas de optimización, como el algoritmo de Optimización por Enjambre de Partículas (PSO), para mejorar la precisión y robustez de los modelos en la detección de deepfakes de audio. Estas arquitecturas híbridas han mostrado una alta efectividad al permitir que los modelos capturen tanto las características espaciales como temporales del audio, lo cual es fundamental para detectar manipulación en tiempo real y con una precisión elevada (Al-Adwan et al., 2024).

### 2.2.4. Estudios Previos en Detección de Deepfakes

Varios estudios han analizado el desempeño de diferentes técnicas para detectar deepfakes. Por ejemplo, Yi et al. (2023) presentan una revisión de métodos de detección de deepfakes de audio, destacando cómo las características espectrales del habla (como el análisis de Mel-spectrogramas) pueden ser útiles para identificar anomalías en el audio. Además, los métodos basados en prosodia han demostrado ser efectivos para identificar patrones irregulares en deepfakes de audio, ya que los modelos generativos tienden a producir una prosodia ”plana.º con variaciones artificiales.

Otro enfoque reciente es el uso de aprendizaje auto-supervisado, en el cual el modelo aprende a identificar patrones de autenticidad sin necesidad de una gran cantidad de datos etiquetados. Lanzino et al. (2024) proponen un método de auto-supervisión que permite a los modelos capturar patrones de voz natural y detectar deepfakes con mayor precisión. Este método es particularmente útil en escenarios donde los datos etiquetados son limitados o difíciles de obtener, y ha demostrado ser prometedor para mejorar la generalización de los modelos en la detección de deepfakes.

### 2.2.5. Desafíos y Limitaciones en la Detección de Deepfakes de Audio

A pesar de los avances, los sistemas de detección de deepfakes enfrentan varios desafíos. Uno de los principales problemas es la capacidad de generalización de los modelos, es decir, su habilidad para detectar deepfakes creados con nuevas técnicas que no han sido vistas durante el entrenamiento. Los modelos de detección tienden a especializarse en ciertos tipos de manipulación, lo que los hace vulnerables a ataques desconocidos. Esto plantea la necesidad de desarrollar modelos más robustos que puedan adaptarse a nuevas técnicas de generación de deepfakes (Heidari et al., 2023; Oorloff et al., 2024).

Otro desafío importante es la interpretabilidad de los modelos. Dado que muchos mo-

delos de detección de deepfakes son complejos y funcionan como çajas negras", es difícil para los investigadores y los usuarios entender cómo el modelo toma decisiones. La interpretabilidad es crucial en aplicaciones de seguridad y en contextos legales, donde es necesario explicar por qué un audio se clasifica como falso o auténtico (Yi et al., 2023).

## 2.3. Marco Conceptual

### 2.3.1. Inteligencia Artificial (IA) y Aprendizaje Profundo

La inteligencia artificial (IA) ha transformado la forma en que interactuamos con la tecnología, permitiendo que las máquinas realicen tareas que normalmente requieren habilidades humanas, como la interpretación del lenguaje, la visión y el reconocimiento de patrones. El aprendizaje profundo es una subdisciplina clave dentro de la IA que utiliza redes neuronales profundas (DNN) para analizar grandes volúmenes de datos y extraer patrones complejos (Goodfellow, Bengio, & Courville, 2017). En lugar de depender de reglas explícitas programadas, las DNN aprenden directamente de los datos, permitiéndoles realizar predicciones precisas al procesar información no estructurada, como el audio. Este tipo de aprendizaje es especialmente útil en la detección de deepfakes, ya que permite que los modelos distingan entre voces reales y generadas artificialmente al identificar discrepancias en el tono, timbre y otros rasgos característicos de la voz humana (Mitchell, 2019).

Las redes neuronales profundas constan de múltiples capas, cada una de las cuales realiza una serie de cálculos para extraer características de mayor nivel de abstracción a partir de los datos de entrada. La representación jerárquica que logran estas redes permite identificar patrones complejos en señales de voz, tales como las fluctuaciones en el tono o las transiciones suaves entre fonemas, que son esenciales para reconocer la autenticidad de una grabación de voz. En el caso de los deepfakes de audio, el aprendizaje profundo facilita la detección de irregularidades en los patrones de voz generados artificialmente, lo cual es crucial para prevenir fraudes y proteger la identidad en el entorno digital (Mitchell, 2019).

### 2.3.2. Procesamiento de Lenguaje Natural (PLN) y Procesamiento de Señales de Voz

El procesamiento de lenguaje natural (PLN) se centra en la interacción entre computadoras y el lenguaje humano, permitiendo que las máquinas interpreten, generen y respondan al lenguaje de manera coherente. Cuando se trata de detección de deepfakes en audio, el PLN

se extiende al análisis del habla, combinándose con el procesamiento de señales de voz, que transforma las ondas de sonido en datos digitales para su posterior análisis (Rao & McMahan, 2019). A través de técnicas de PLN, es posible descomponer la señal de audio en sus elementos constitutivos (frecuencias, duraciones, intensidades) y analizar características como la entonación y el ritmo, que son fundamentales para identificar las particularidades de una voz auténtica (Rabiner & Schafer, 2007).

La señal de voz se caracteriza por elementos acústicos que incluyen el tono, el timbre y la prosodia, cada uno de los cuales aporta información sobre el hablante y la autenticidad de su discurso. En el contexto de la detección de deepfakes, estos elementos permiten establecer patrones que las voces sintéticas encuentran difícil replicar con precisión. Por ejemplo, el tono y la prosodia pueden variar según el contexto y las emociones del hablante, una característica que suele perderse en las voces generadas artificialmente (Rabiner & Schafer, 2007).

### 2.3.3. Deepfakes de Audio

Los deepfakes de audio representan una de las aplicaciones más avanzadas y potencialmente peligrosas del aprendizaje profundo. Estas falsificaciones de voz se logran mediante modelos generativos, como los modelos de redes neuronales recurrentes (RNN) o las redes de confrontación generativa (GAN), que pueden sintetizar voces convincentes a partir de muestras de audio de una persona específica. El resultado es un audio que imita la voz y el estilo de habla de una persona real, lo que puede ser usado en actividades fraudulentas, como la suplantación de identidad en llamadas telefónicas o la creación de grabaciones de voz falsas (Gomes-Gonçalves, 2022).

La creación de deepfakes de audio plantea desafíos éticos y de seguridad, ya que permite la manipulación de información de manera convincente, desafiando la autenticidad de las comunicaciones y poniendo en riesgo la seguridad de personas y organizaciones. Debido a que las voces humanas tienen características únicas y sutiles, los deepfakes en audio aún presentan fallos al intentar imitar todos los detalles acústicos, como los patrones de ritmo y la entonación, lo cual se convierte en una ventaja para los sistemas de detección basados en IA (Mitchell, 2019).

### 2.3.4. Variables de Estudio para la Detección de Deepfakes en Audio

En la detección de deepfakes de audio, se emplean variables acústicas clave que permiten analizar las diferencias entre una grabación genuina y una generada artificialmente. Cada

una de estas variables proporciona información única sobre las características de la voz y son esenciales para identificar posibles manipulaciones.

#### 2.3.4.1. Tono de Voz (Pitch)

El tono es la frecuencia fundamental de la voz y se relaciona con la percepción de si un sonido es grave o agudo. En la voz humana, el tono varía naturalmente según factores emocionales y fisiológicos, creando un patrón que es difícil de replicar exactamente en deepfakes. Los sistemas de detección utilizan el análisis de tono para observar irregularidades en estas variaciones, que pueden indicar una manipulación de la voz (Rabiner & Schafer, 2007).

#### 2.3.4.2. Timbre de Voz

El timbre se refiere a la çalidad.º çolor"de la voz, influenciado por las características físicas del tracto vocal del hablante. Cada persona tiene un timbre único debido a diferencias en la estructura de sus cuerdas vocales y en la resonancia del tracto vocal. El timbre es especialmente importante en la detección de deepfakes, ya que los modelos de generación de voz suelen tener dificultades para imitar esta característica distintiva de forma precisa, lo que permite detectar falsificaciones a partir de anomalías en el timbre (Jurafsky & Martin, 2024).

#### 2.3.4.3. Duración y Ritmo del Habla

La duración de los sonidos y el ritmo general del habla son patrones consistentes en cada persona, lo que significa que un cambio en el ritmo puede indicar una manipulación. Las voces generadas tienden a presentar una cadencia que puede sonar artificial o robotizada", especialmente cuando no logran replicar pausas o variaciones naturales en la duración de las sílabas y palabras (Rao & McMahan, 2019).

#### 2.3.4.4. Mel-Spectrograma

El mel-espectrograma es una herramienta visual que representa la energía de una señal de audio en diferentes frecuencias y permite identificar patrones espectrales característicos. Las voces reales suelen tener un espectrograma mel con transiciones suaves entre frecuencias, mientras que los deepfakes pueden mostrar patrones espectrales diferentes o artefactos debido a la síntesis. Analizar el mel-espectrograma permite a los modelos de IA detectar las irregularidades propias de las voces generadas (Rabiner & Schafer, 2007).

#### 2.3.4.5. Prosodia (Entonación y Énfasis)

La prosodia, que incluye variaciones en la entonación y el énfasis, proporciona un ritmo natural y una fluidez al habla humana. Estas variaciones son difíciles de imitar en las voces sintéticas, que tienden a tener una prosodia ”plana.º inconsistente. Los sistemas de detección de deepfakes analizan la prosodia para identificar patrones irregulares que puedan indicar una generación artificial (Goodfellow et al., 2017).

#### 2.3.4.6. Articulación y Transiciones entre Fonemas

La articulación de los sonidos del habla y las transiciones entre ellos son suaves en una voz humana. En los deepfakes, las transiciones entre fonemas pueden ser bruscas o presentar errores de articulación, debido a las limitaciones de los modelos de generación de voz. Este análisis permite detectar cuando una voz carece de la fluidez natural de la pronunciación humana, ayudando a identificar una grabación manipulada (Jurafsky & Martin, 2024).

# Capítulo 3

# METODOLOGÍA DE LA INVESTIGACIÓN

## 3.1. Diseño de la investigación

En esta sección del documento se explicará cual es el diseño, el tipo y el enfoque del trabajo de investigación, así como también la población y la muestra.

### 3.1.1. Enfoque de la investigación

El enfoque de esta investigación es cuantitativo y experimental, dirigida a mejorar la detección de deepfakes en audio en español. A través del uso de redes neuronales profundas, este estudio explora específicamente las características acústicas del audio (como tono, timbre, patrones de voz y otros atributos prosódicos) para diferenciar entre audios genuinos y manipulados. Este enfoque se basa en la implementación y evaluación de modelos de aprendizaje profundo y técnicas de procesamiento de señales, buscando no solo la precisión en la detección sino también la adaptación a las particularidades del idioma español y al contexto peruano, donde los casos de fraude por deepfakes están en aumento.

### 3.1.2. Alcance de la investigación:

El alcance de esta investigación incluye el desarrollo y validación de un modelo de detección de deepfakes de audio optimizado para audios en español. La investigación cubrirá tanto la identificación de los atributos acústicos más relevantes en la detección de deepfakes

en español como la construcción de una base de datos experimental que incorpore muestras de audio adaptadas al contexto peruano. Además, este estudio analizará la efectividad del modelo en varios escenarios acústicos (diferentes niveles de ruido de fondo y calidad de grabación), ampliando su aplicabilidad en situaciones reales de verificación de identidad, seguridad digital y prevención de fraudes.

### 3.1.3. EDiseño de la investigación:

El diseño de esta investigación es experimental y longitudinal. Involucra la implementación de un sistema de detección de deepfakes que será entrenado, validado y probado en múltiples etapas. Primero, se seleccionarán y procesarán muestras de audio de deepfakes y genuinas en español para obtener un conjunto de datos representativo. Luego, se aplicarán diferentes arquitecturas de redes neuronales profundas, evaluando sus precisiones y ajustando parámetros clave para optimizar el rendimiento. El diseño incluye comparativas en la precisión del modelo y pruebas de robustez frente a variaciones en las características del audio, utilizando métricas como la tasa de error igual (EER) y el F1 score. Finalmente, el estudio evaluará la generalización del modelo frente a datos no observados previamente, asegurando su aplicabilidad en contextos reales y su utilidad para mitigar el fraude.

## 3.2. Población y muestra

La población de estudio para esta investigación estará compuesta por todos los ciudadanos peruanos que potencialmente podrían ser víctimas de fraudes a través del uso de deepfake en español. Esta población incluirá individuos de diversas regiones del país, considerando tanto áreas urbanas como rurales, lo cual permitirá obtener una visión amplia y representativa de las distintas realidades socioeconómicas y tecnológicas presentes en el Perú. Se pondrá especial énfasis en aquellos que se encuentren expuestos a situaciones de riesgo en plataformas de comunicación digital, redes sociales y medios de comunicación tradicionales, ya que estos medios son los principales canales de distribución del contenido manipulado mediante deepfake.

Además, se considerarán distintos perfiles demográficos, incluyendo personas de diferentes niveles socioeconómicos, rangos de edad y grados de acceso a tecnologías digitales. Esto permitirá analizar cómo factores como la educación, el ingreso económico y el acceso a la tecnología influyen en la susceptibilidad frente a fraudes mediante deepfake. Incluir a individuos con distintos niveles de alfabetización digital permitirá evaluar si el conocimiento y manejo de tecnologías de la información y comunicación influyen en la capacidad de identificar y resistir

posibles fraudes.

Según Hernández Sampieri et al. (2018), la población se define como .[el] conjunto de todos los casos que concuerdan con una serie de especificaciones". En este caso, los çasos"se refieren a personas que cumplen con características como el idioma (español) y la exposición a contextos vulnerables al fraude digital en el Perú. Definir con claridad la población permite precisar los objetivos del estudio y delimitar los criterios de inclusión y exclusión, lo cual es crucial para la validez de la investigación. Además, contar con una población bien delimitada facilita la representatividad de los resultados y asegura que las conclusiones obtenidas puedan aplicarse al contexto general peruano, proporcionando datos relevantes para la prevención y mitigación de fraudes mediante deepfake.

La muestra de este estudio se seleccionará utilizando un muestreo no probabilístico intencional, debido a la naturaleza específica de la población objetivo y a las características del fenómeno a estudiar. La muestra estará conformada por aproximadamente 500 individuos de distintas regiones del Perú, quienes serán seleccionados por cumplir con los criterios de exposición a riesgos de fraude mediante deepfake, así como por su participación activa en plataformas digitales y redes sociales. Esta muestra buscará representar una diversidad de contextos socioeconómicos y tecnológicos, lo cual permitirá evaluar cómo diferentes factores influyen en la vulnerabilidad frente a estos fraudes. El tamaño de la muestra se ha determinado para asegurar una representatividad adecuada y permitir un análisis significativo de las variables estudiadas, garantizando al mismo tiempo la viabilidad operativa del estudio.

## 3.3. Operacionalización de Variables

### 3.3.1. Variable dependiente

#### 3.3.1.1. Detección de audio deepfake

- Definición conceptual: Proceso de identificación de audios generados mediante técnicas de síntesis y manipulación de voz con el objetivo de suplantar la identidad de una persona o de alterar el contenido del mensaje.
- Definición operacional: Medida binaria (1 = deepfake, 0 = no deepfake) determinada por el modelo de redes neuronales profundas diseñado para analizar y clasificar el audio como genuino o falso. La detección se realiza a partir de la salida del modelo, utilizando métricas de precisión, sensibilidad y especificidad para evaluar el desempeño.

## 3.3.2. Variables independientes

#### 3.3.2.1. Patrones de voz

- Definición conceptual: Proceso de identificación de audios generados mediante técnica-Definición conceptual: Características recurrentes del habla que incluyen aspectos como el ritmo, la cadencia y los matices propios de la voz de una persona.
- Definición operacional: Análisis cuantitativo de patrones acústicos extraídos de señales de audio mediante la extracción de características como Mel-spectrogramas y coeficientes cepstrales de frecuencia Mel (MFCC).

#### 3.3.2.2. Frecuencia fundamental (pitch)

- Definición conceptual: Frecuencia básica de vibración de las cuerdas vocales durante la producción de la voz, que se percibe como la altura tonal.
- Definición operacional: Medición en Hz de la frecuencia fundamental mediante algoritmos de análisis de señal (autocorrelación) para determinar la variación del pitch a lo largo del tiempo en cada fragmento de audio.

#### 3.3.2.3. Duración y ritmo del habla

- Definición conceptual: Duración de los sonidos emitidos y la velocidad a la que se emiten las palabras o sílabas.
- Definición operacional: Medición del tiempo promedio (en segundos) que toma cada sílaba y cálculo del ritmo del habla a partir de la distribución temporal de las pausas y la velocidad promedio del discurso.

#### 3.3.2.4. Tono de voz

- Definición conceptual: Percepción del sonido que permite identificar si es grave o agudo, relacionado con la modulación del pitch.
- Definición operacional: Evaluación subjetiva a partir del análisis de la variación del pitch, utilizando técnicas de cuantificación estadística de la frecuencia fundamental.

#### 3.3.2.5. Timbre de voz

- Definición conceptual: Calidad del sonido que distingue una voz de otra, incluso si tienen el mismo pitch y volumen.
- Definición operacional: Análisis de la señal mediante la extracción de armónicos y características espectrales (formantes) que permiten diferenciar la voz individual.

#### 3.3.2.6. Formantes

- Definición conceptual: Picos de resonancia en el espectro de la voz que permiten caracterizar los sonidos del habla.
- Definición operacional: Identificación y medición de las frecuencias de los principales formantes (F1, F2, F3) mediante técnicas de análisis espectral.

#### 3.3.2.7. Prosodia

- Definición conceptual: Aspectos suprasegmentales del habla, como el acento, la entonación y el ritmo.
- Definición operacional: Análisis de la variación de la frecuencia fundamental y la amplitud para medir patrones prosódicos y su correspondencia con el contenido del habla.

#### 3.3.2.8. Articulación

- Definición conceptual: Movimiento y posición de los órganos articulatorios (labios, lengua, paladar) al momento de producir sonidos.
- Definición operacional: Evaluación indirecta a partir del análisis de transiciones suaves y precisas entre fonemas, utilizando el espectrograma para identificar inconsistencias.

#### 3.3.2.9. Transiciones entre fonemas

- Definición conceptual: Fluidez y continuidad en el paso de un fonema a otro durante el habla.
- Definición operacional: Análisis espectral que identifica la calidad de la transición fonética, buscando indicios de discontinuidades o anomalías entre fonemas.

#### 3.3.2.10. Ruido de fondo

- Definición conceptual: Sonidos adicionales presentes durante la grabación de la voz que no son parte de la emisión vocal intencional.
- Definición operacional: Detección y cuantificación del nivel de ruido de fondo (en dB) utilizando técnicas de reducción de ruido y análisis espectral para evaluar la influencia del entorno.

## 3.4. Técnicas de recolección de datos

Para la recolección de datos en esta investigación sobre la detección de audio deepfake en español utilizando redes neuronales profundas, se considerarán varias técnicas de recolección de datos, principalmente enfocadas en la obtención de muestras de audio que sean representativas del fenómeno deepfake y del habla natural.

### 3.4.1. Grabaciones Directas de Voz

#### 3.4.1.1. Participantes Voluntarios

Se recopilarán datos grabados directamente a voluntarios hablando en español. Para esto, se realizarán sesiones de grabación bajo condiciones controladas, donde los participantes hablen sobre ciertos temas específicos.

#### 3.4.1.2. Variación de Condiciones de Grabación

Se incluirán también voces en diferentes ambientes para así evaluar el ruido de fondo en la detección de deepfake.

### 3.4.2. Recolección de Audios de Plataformas de Comunicación

#### 3.4.2.1. Redes sociales y plataformas de mensajería

Con el consentimiento adecuado y cumpliendo con las normativas éticas y de privacidad, se podrá recopilar audios de plataformas como WhatsApp, Telegram, Facebook,etc.

#### 3.4.2.2. Audios de YouTube, Podcasts u otras plataformas de creación de contenido

Se tomarán audios de plataformas públicas en los que se encontrarán audios de diversas temáticas y acentos propios de las diferentes regiones geográficas.

### 3.4.3. Generación de Audio Deepfake

#### 3.4.3.1. Generación con Modelos TTS y de Conversión de Voz

Se pueden recolectar datos generando audios deepfake utilizando herramientas de Text-to-Speech (TTS) como Tacotron, WaveNet, FastSpeech, o técnicas de conversión de voz como Voice Conversion. Se pueden crear audios falsos que emulen las voces de ciertos individuos y luego comparar estos con audios genuinos. Esto sería parte de la generación controlada de audios para su uso como datos de entrenamiento.

#### 3.4.3.2. Sistemas de Generación Basados en IA

Se generarán audios falsos utilizando redes neuronales generativas como GANs (Generative Adversarial Networks) o variaciones como StyleGAN. También se podrían usar sistemas como DeepVoice, que están específicamente diseñados para manipular el audio.

## 3.5. Técnicas para el procesamiento y análisis de la información

### 3.5.1. Extracción de Características Acústicas

A continuación se detallan las técnicas fundamentales para extraer la información relevante de la señal de audio que servirá como insumo para el entrenamiento de los modelos de detección.

- MFCC (Coeficientes Cepstrales en Frecuencia Mel): Los MFCC se utilizan para obtener una representación espectral de la señal de audio que refleja las características del tracto vocal. Esta técnica es útil para analizar el timbre y distinguir entre diferentes voces (Zhang et al., 2017).

- Formantes: Los formantes son picos de resonancia en el espectro de la voz que caracterizan a los diferentes sonidos del habla. Medir las frecuencias de los principales formantes (F1, F2, F3) permite detectar diferencias en la articulación, útiles para identificar audios deepfake (Jacewicz & Fox, 2013).

- Formantes: Los formantes son picos de resonancia en el espectro de la voz que caracterizan a los diferentes sonidos del habla. Medir las frecuencias de los principales formantes (F1, F2, F3) permite detectar diferencias en la articulación, útiles para identificar audios deepfake (Jacewicz & Fox, 2013).

- Prosodia: La prosodia abarca aspectos como la entonación, el ritmo y la duración del habla. Estos factores pueden ser útiles para detectar anomalías típicas de audios deepfake, ya que la generación artificial de voz puede tener patrones prosódicos que no coinciden con el habla humana natural (Rosenberg, 2012).

- Frecuencia Fundamental (Pitch): Analizar la frecuencia fundamental del audio permite evaluar la estabilidad y consistencia del tono de voz. Los cambios abruptos o inconsistentes pueden ser un indicativo de manipulación (Boersma & Weenink, 2019).

- Frecuencia Fundamental (Pitch): Analizar la frecuencia fundamental del audio permite evaluar la estabilidad y consistencia del tono de voz. Los cambios abruptos o inconsistentes pueden ser un indicativo de manipulación (Boersma & Weenink, 2019).

- Duración y Ritmo del Habla: La duración de fonemas y el ritmo general del habla son útiles para detectar diferencias entre audios genuinos y deepfake. Los modelos generativos suelen producir ritmos poco naturales o incoherencias en las transiciones entre fonemas (Cummins, 2018).

- Duración y Ritmo del Habla: La duración de fonemas y el ritmo general del habla son útiles para detectar diferencias entre audios genuinos y deepfake. Los modelos generativos suelen producir ritmos poco naturales o incoherencias en las transiciones entre fonemas (Cummins, 2018).

- Ruido de Fondo: Analizar el ruido de fondo, así como la reducción de ruido, puede ayudar a identificar inconsistencias que podrían indicar un audio falsificado. El ruido generado por deepfakes puede no ser coherente con el contenido del audio o con el ambiente en el que fue grabado (Taal et al., 2011).

### 3.5.2. Transformadas y Representaciones Espectrales

Para convertir la señal de audio en representaciones que sean más fáciles de analizar y útiles para los modelos.

- Transformada de Fourier (FFT): La Transformada Rápida de Fourier convierte la señal del dominio temporal al dominio frecuencial, permitiendo analizar cómo se distribuyen las frecuencias. Esto ayuda a detectar patrones de manipulación en las frecuencias del audio (Smith, 2011).
- Mel-Spectrograma: El Mel-spectrograma es una representación que muestra la intensidad de las frecuencias en función del tiempo. Es una herramienta visual poderosa para analizar la energía y detectar anomalías que pueden ser comunes en audios deepfake (Broughton & Donovan, 2012).

### 3.5.3. Técnicas de Análisis de la Señal de Audio

Estas técnicas permiten una mejor comprensión de la señal y facilitan la extracción de información relevante.

- Autocorrelación para Extracción de Pitch: Esta técnica se utiliza para medir la frecuencia fundamental (pitch), lo cual es útil para identificar inconsistencias en el tono de voz (Boersma & Weenink, 2019).
- Análisis de Energía del Habla: Medir la energía o amplitud de la señal a lo largo del tiempo permite identificar variaciones inesperadas en la intensidad del habla. Los audios deepfake pueden mostrar fluctuaciones de energía poco naturales (Mitra et al., 2016).

### 3.5.4. Modelos de Machine Learning y Deep Learning

Para la clasificación de audios como genuinos o falsos, los modelos de aprendizaje son el componente principal.

- Redes Neuronales Convolucionales (CNN): Las CNN son útiles para analizar representaciones espectrales (como Mel-spectrogramas) y extraer patrones que diferencien entre audios genuinos y deepfake. Capturan características espaciales de las representaciones espectrales (Goodfellow et al., 2016).

- Redes Neuronales Recurrentes (RNN) y LSTM (Long Short-Term Memory): Las RNN y variantes como LSTM se utilizan para capturar dependencias temporales en la señal de audio, lo cual es útil para modelar la dinámica del habla. Estas redes ayudan a identificar anomalías en la continuidad del flujo del habla (Hochreiter & Schmidhuber, 2019).
- Transformers: Los modelos basados en transformers también pueden aplicarse al análisis de audio para capturar patrones complejos de larga distancia, lo cual es útil en el caso de audios con contextos amplios (Vaswani et al., 2017).

### 3.5.5. Algoritmos de Machine Learning Tradicionales

Pueden ser utilizados como métodos adicionales para la clasificación o en combinación con redes neuronales profundas.

- SVM (Support Vector Machine): Las máquinas de vectores soporte son útiles para la clasificación binaria de audios utilizando las características acústicas extraídas. Se pueden emplear como un paso de preclasificación o como un método independiente (Cortes & Vapnik, 2010).
- Random Forest: Un algoritmo basado en árboles de decisión que permite combinar múltiples características del audio para realizar la clasificación. Este método es útil para manejar datos de alta dimensionalidad y puede proporcionar interpretaciones útiles sobre la importancia de cada característica (Breiman, 2011).

### 3.5.6. Técnicas de Preprocesamiento

Antes de utilizar cualquier modelo de análisis, es necesario realizar ciertas etapas de preprocesamiento.

- Normalización de la Señal de Audio: Consiste en estandarizar la amplitud del audio para asegurar que todas las señales tengan un nivel de volumen similar y que el modelo no se vea influenciado por diferencias en la intensidad (Wang & Brown, 2018).
- Segmentación del Audio: Dividir los audios en segmentos más pequeños puede facilitar la identificación de patrones. Esto es especialmente útil cuando los audios tienen duraciones largas, ya que facilita el entrenamiento y mejora la capacidad del modelo para detectar patrones en fragmentos más cortos (Huang et al., 2014).

## 3.6. Cronograma de actividades y presupuesto

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

| Item | Quantity |
|---|---|
| Widgets | 42 |
| Gadgets | 13 |

**Tabla 3.1:** An example table.

# Capítulo 4

# DESARROLLO DEL EXPERIMENTO

## 4.1. X

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn? Kjift ”not at all!...

## 4.2. Y

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

| Item | Quantity |
|---|---|
| Widgets | 42 |
| Gadgets | 13 |

**Tabla 4.1:** An example table.

## 4.3. Z

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

El paper es citado y el otro paper .

# Capítulo 5

# ANÁLISIS Y DISCUSIÓN DE RESULTADOS

## 5.1. X

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn? Kjift ”not at all!...

## 5.2. Y

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

| Item | Quantity |
|---|---|
| Widgets | 42 |
| Gadgets | 13 |

**Tabla 5.1:** An example table.

## 5.3. Z

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

# Capítulo 6

# CONCLUSIONES Y RECOMENDACIONES

## 6.1. Conclusiones

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn? Kjift ”not at all!...

## 6.2. Recomendaciones

Nisi porta lorem mollis aliquam ut porttitor leo. Aenean pharetra magna ac placerat vestibulum. Est placerat in egestas erat imperdiet sed euismod. Velit euismod in pellentesque massa placerat. Enim praesent elementum facilisis leo vel fringilla. Ante in nibh mauris cursus mattis molestie a iaculis. Erat pellentesque adipiscing commodo elit at imperdiet dui accumsan sit. Porttitor lacus luctus accumsan tortor posuere ac ut. Tortor at auctor urna nunc id. A iaculis at erat pellentesque adipiscing commodo elit.

# Anexos

# Anexos A

# Anexo I: Matriz de Consistencia

| PROBLEMAS | OBJETIVOS | HIPÓTESIS |
|---|---|---|
| Problema General | Objetivo General | Hipótesis General |
| El incremento del fraude en Perú mediante el uso de tecnologías deepfake de audio ha evidenciado la falta de herramientas adecuadas para detectar estos fraudes, especialmente en español. Las técnicas actuales no logran identificar eficazmente las características acústicas del español, como el tono de voz, timbre de voz, patrones de voz, frecuencia fundamental (pitch), duración y ritmo del habla, formantes, nivel de energía del habla (intensidad), ruidos de fondo, prosodia, articulación y transiciones entre fonemas, lo que facilita la suplantación de identidad y el fraude en las comunicaciones personales y empresariales. | Desarrollar un modelo basado en redes neuronales profundas que permita detectar deepfakes de audio en español mediante el análisis de variables clave como tono de voz, timbre de voz, patrones de voz, frecuencia fundamental, duración y ritmo del habla, formantes, nivel de energía del habla, ruidos de fondo, prosodia, articulación y transiciones entre fonemas, mejorando la precisión en la identificación de audios manipulados para mitigar fraudes por suplantación de identidad en Perú. | El uso de un modelo basado en redes neuronales profundas que analice las variables acústicas clave como tono de voz, timbre de voz, patrones de voz, frecuencia fundamental, duración y ritmo del habla, formantes, nivel de energía del habla, ruidos de fondo, prosodia, articulación y transiciones entre fonemas mejora significativamente la precisión en la detección de deepfakes de audio en español, reduciendo el riesgo de fraudes por suplantación de identidad en Perú. |
| Problemas Específicos | Objetivos Específicos | Hipótesis Específicas |
| La falta de un dataset en español que incluya variaciones regionales y voces manipuladas dificulta el entrenamiento de modelos de redes neuronales profundas para detectar deepfakes de audio en español, debido a las diferencias en patrones de voz, frecuencia fundamental (pitch) y ritmo del habla. | Desarrollar un dataset específico en español, con variaciones regionales y voces manipuladas, para entrenar un modelo de redes neuronales profundas que detecte deepfakes de audio | La creación de un dataset en español que incluya variaciones regionales y voces manipuladas mejorará significativamente la capacidad de las redes neuronales profundas para detectar deepfakes de audio en este idioma |
| Las técnicas actuales no logran detectar las variaciones en el tono de voz, timbre de voz y formantes en audios en español, lo que disminuye la precisión en la identificación de audios manipulados. | Implementar un modelo de redes neuronales profundas que analice el tono de voz, timbre de voz y formantes para mejorar la precisión en la detección de deepfakes de audio en español. | El análisis del tono de voz, timbre de voz y formantes mediante redes neuronales profundas aumentará la precisión en la detección de deepfakes de audio en español. |
| Los fraudes por suplantación de identidad mediante deepfakes de audio en Perú son difíciles de detectar con las técnicas actua- | Evaluar la eficacia del modelo de redes neuronales profundas en la detección de deepfakes de audio | El modelo de redes neuronales profundas será más efectivo en la detección de deepfakes en contextos de fraude en |

# Anexos B

# Anexo II: Resumen de Papers investigados

| Tipo | N° | Título | Autor | Año | País | Fuente |
|---|---|---|---|---|---|---|
| Problema | 1 | Copper price estimation using bat algorithm | Dehghani Bogdanovic | 2018 | United Kingdom | Resources Policy |
| | 2 | Alternative techniques for forecasting mineral commodity prices | Cortez, Saydam, Coulton, Sammut | 2018 | Netherlands | International Journal of Mining Science and Technology |
| Propuesta | 3 | Prediction of the crude oil price thanks to natural language processing applied to newspapers | Trastour, Genin, Morlot | 2016 | USA | Standfort University ML repository |
| | 4 | Stock Price Prediction Using Deep Learning | Tipirisetty | 2018 | USA | Master's Theses San Jose State University |
| | 5 | Deep Learning for Stock Prediction Using Numerical and Textual Information | Akita, R., Yoshihara, A., Matsubara, T., Uehara, K. | 2016 | USA | 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS) |
| Técnica | 6 | Stock Prices Prediction using the Title of Newspaper Articles with Korean Natural Language Processing | Yun, Sim, Seok | 2019 | Japan | 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC) |
| | 7 | A Method of Optimizing LDA Result Purity Based on Semantic Similarity | Jingrui, Z., Qinglin, W., Yu, L., Yuan, L. | 2017 | China | 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC) |
| | 8 | Qualitative Stock Market Predicting with Common Knowledge Based Nature Language Processing: A Unified View and Procedure | Rao, D., Deng, F., Jiang, Z., Zhao, G. | 2015 | USA | 2015 7th International Conference on Intelligent Human-Machine Systems and Cybernetics |
| | 9 | Fuzzy Bag-of-Words Model for Document Representation | Zhao, R., Mao, K. | 2018 | USA | IEEE Transactions on Fuzzy Systems ( Volume: 26 , Issue: 2 , April 2018 ) |

**Tabla B.1:** Cuadro Resumen de Papers investigados. Fuente: Elaboración propia