Desarrollo de un algoritmo detector de engaños mediante la utilización del método Deep Learning de inteligencia artificial

Chango Salas, Jorge Alejandro

Departamento de Eléctrica, Electrónica y Telecomunicaciones

Carrera de Ingeniería en Electrónica y Telecomunicaciones

Trabajo de titulación, previo a la obtención de Ingeniero en Electrónica y Telecomunicaciones

Ing. Larco Bravo, Julio César

26 de Noviembre de 2021

Informe de originalidad

NOMBRE DEL CURSO
Titulación

Firmado electrónicamente por:
JULIO CESAR LARCO BRAVO

NOMBRE DEL ALUMNO
JORGE ALEJANDRO CHANGO SALAS

NOMBRE DEL ARCHIVO
JORGE ALEJANDRO CHANGO SALAS - Trabajo de titulación Jorge Chango

SE HA CREADO EL INFORME
8 dic 2021

Resumen

Fragmentos marcados	14	2 %
Fragmentos citados o entrecomillados	11	2 %

Coincidencias de la Web

espe.edu.ec	16	3 %
uss.edu.pe	5	0,7 %
passeidireto.com	1	0,2 %
scribd.com	1	0,2 %
upm.es	1	0,1 %
gobiernodecanarias.org	1	0,1 %

1 de 25 fragmentos

Fragmento del alumno MARCADO

Trabajo de titulación, previo a la obtención de Ingeniero en Electrónica y Telecomunicaciones Ing. Larco Bravo, Julio César

Mejor coincidencia en la Web

Trabajo de titulación, previo a la obtención del título **de Ingeniero en Electrónica y Telecomunicaciones Ing**. Bernal Oñate, Carlos Paúl, MSc.

1 Desarrollo de un algoritmo de reconocimiento de emociones
... http://repositorio.espe.edu.ec/bitstream/21000/25641/1/T-ESPE-044698.pdf

2 de 25 fragmentos

Fragmento del alumno CITADO

…Machine Learning, en una investigación de Bravo Santiago denominada "**Implementación de un sistema de reconocimiento automático de engaños mediante el análisis de la señal de la voz**

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Desarrollo de un algoritmo detector de engaños mediante la utilización del método Deep Learning de inteligencia artificial**" fue realizado por el señor **Chango Salas, Jorge Alejandro** el cual ha sido revisado y analizado en su totalidad por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 25 de Noviembre 2021

Firma:

Firmado electrónicamente por:
JULIO CESAR LARCO BRAVO

.....

Ing. Larco Bravo, Julio César

C. C. 1710638808

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

RESPONSABILIDAD DE AUTORÍA

Yo, **Chango Salas, Jorge Alejandro**, con cédula de ciudadanía nº 1717575383, declaro que el contenido, ideas y criterios del trabajo de titulación: **"Desarrollo de un algoritmo detector de engaños mediante la utilización del método Deep Learning de inteligencia artificial"** es de mi autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 25 de Noviembre 2021

Firma

Chango Salas, Jorge Alejandro

C.C.: 1717575383

DEPARTAMENTO DE ELÉCTRICA, ELECTRÓNICA Y TELECOMUNICACIONES

CARRERA DE INGENIERÍA EN ELECTRÓNICA Y TELECOMUNICACIONES

AUTORIZACIÓN DE PUBLICACIÓN

Yo, **Chango Salas, Jorge Alejandro,** con cédula de ciudadania N° 1717575383, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **"Desarrollo de un algoritmo detector de engaños mediante la utilización del método Deep Learning de inteligencia artificial"** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 25 de Noviembre 2021

Firma

Chango Salas, Jorge Alejandro

C.C.: 1717575383

Dedicatoria - Agradecimiento

A la memoria de mi madre, María Elena, quién ha venido cuidando de mí, y siguiendo todos mis pasos desde el plano no terrenal, por quién soy y cumplo la mayor cantidad de sueños y promesas, en su nombre.

A mi padre, Jorge, por su cariño y su apoyo incondicional, por sus consejos que han sabido guiarme hasta donde hoy por hoy me encuentro, y a quién debo mis valores, fortaleza y carácter.

A mis hermanas, Belén y Cristina, a quienes les debo mucho de mi crianza, y de quienes tengo el mejor de los ejemplos de profesionalismo, humanismo y perseverancia, y a quienes admiro por su lucha y dedicación como personal de salud contra el Covid-19. Agradezco siempre su apoyo y guía aún más, en situaciones tan difíciles como las que enfrentamos actualmente.

A mi novia, Cindy, quien ha sido un gran motor y la constante motivación para nunca rendirme en mis objetivos, incluso en los momentos más turbulentos.

A mis amigos, Michael, Santiago, Marlon, Nicole, Alexa, Leonardo, Mika, Kaybett, quienes han sido un brazo más, y una extensión de mi familia, gracias por permitirme aprender más de la vida a su lado, por su aliento y motivación constante.

Agradezco también a mis queridos docentes y amigos con quienes compartimos aulas en medio de muchas risas y conocimiento, pero en especial, a mi director de tesis, quien ha depositado su confianza en mí y ha sabido guiarme en el desarrollo del presente trabajo de titulación.

Jorge Alejandro Chango Salas

Índice de Contenidos

Índice de Tablas

Índice de Figuras

Resumen

En la actualidad estamos regidos por un constante avance dentro del campo tecnológico, por lo que los “engaños” o “estafas” han incrementado paulatinamente, para ello se denota la importancia que tiene la veracidad de información dentro de entornos como en: legislación, publicidad, criminalística, relaciones públicas, trabajo social, psicología; debido a esto, el presente proyecto de investigación, pretende ser una base que será de utilidad para que en futuras investigaciones se considere al Deep Learning como la opción más eficiente para detección de engaños. En el presente proyecto investigativo se realizó el análisis del estrés en la voz a través de la extracción de características de la misma, utilizando para dicho propósito Matlab®. Para determinar un engaño, se consideran las características propias del habla “normal” comparado con las mismas bajo “estrés”. La investigación fue aplicada a una base de datos de 94 señales de audio grabadas entre hombres y mujeres que se presenta de manera pública denominada, “RAVDESS”, mediante las cuales se implementa el reconocimiento automático de verdades y engaños a través del algoritmo de entrenamiento de red neuronal convolucional de Deep Learning, mostrando así, que las características de voz otorgan información necesaria para evaluar o clasificar un engaño de una verdad, obteniendo así, porcentajes de especificidad y sensibilidad adecuados para un análisis confiable de dichos resultados.

PALABRAS CLAVE:

- **DEEP LEARNING**
- **RED NEURONAL CONVOLUCIONAL**
- **DETECCIÓN DE ENGAÑOS**

Abstract

At present we are governed by a constant advance within the technological field, so that "deceptions" or "scams" have gradually increased, for this the importance of the veracity of information is denoted within environments such as: legislation, advertising, criminology, public relations, social work, psychology; For that reason, this project will be a useful base for future research to consider Deep Learning as the most efficient option for detecting deception. In the present research project, the stress analysis in the voice was carried out through the extraction of its characteristics, using Matlab® for this purpose. To determine a deception, the characteristics of "normal" speech are considered compared to those under "stress". The research was applied to a database of 94 audio signals recorded between men and women that is presented in a public way called, “RAVDESS”, by means of which the automatic recognition of truths and deceptions is implemented through the algorithm of training of Deep Learning convolutional neural network, thus showing that the voice characteristics provide the necessary information to evaluate or classify a deception of a truth, thus obtaining adequate specificity and sensitivity percentages for a reliable analysis of said results.

KEYWORDS:

- **DEEP LEARNING**
- **CONVOLUTIONAL NEURAL NETWORK**
- **DECEPTION DETECTION**

Capítulo I

Introducción

Introducción del Proyecto de Investigación

Dentro del estudio galardonado con el premio de investigación UANIL, (Campos, Lopez, & Morales, 2013) afirman que la capacidad de mentir de una persona en promedio es de 1 a 2 veces por día, asimilando que prácticamente el engaño o mentira es parte de la vida diaria del ser humano, actualmente la presente investigación propone aplicar un algoritmo por medio de la inteligencia artificial utilizando la grabación de audio para determinar si una persona miente o no, es decir, desarrollar una red neuronal entrenada en Matlab con el método de aprendizaje Deep Learning para obtener por medio de su análisis un cálculo de predicción para comprobar si el individuo está mintiendo o no.

La forma en la que se realizó el trabajo de titulación, es por medio de una base de datos de grabaciones de voz, con el fin de entrenar a las neuronas y enfocarse en el grado de estrés a la que está sometida la persona, y determinar la veracidad de lo que se esté tratando.

La razón de la investigación es poder demostrar el grado de confiabilidad que tiene la utilización de Redes Neuronales Convolucionales (CNN) como parte de los algoritmos de Deep Learning para detectar los engaños por medio de la voz, además comparar la optimización que mantiene respecto a otras metodologías que se utiliza para detectar los engaños. así mismo, el automatizar e innovar esta herramienta de detección de mentiras es fundamental para la sociedad donde la verdad es una variable imprescindible con motivo de, declarar a una persona inocente o culpable, evitar fraudes, salvar a personas del suicidio, determinar si las llamadas de emergencia son veraces, así pues, el algoritmo, tiene múltiples beneficios que darían un aporte significativo para las sociedades

Antecedentes

Por décadas el tipificar un proceso confiable para determinar si una persona miente o dice la verdad ha sido complejo, muchos estudios han implementado variables a identificar como la expresión facial, gestos, el tono de la voz, etc.

En muchos casos, el éxito para determinar la confiabilidad del acto de mentir ha sido solo del 50%, así mismo, una habilidad para detectar el engaño es influenciada por procesos cognitivos, dicho de otra manera, quien miente tiene mayor recurso cognitivo, pero existen pocos estudios respecto a ese tipo de metodología implantada para determinar el engaño, sea el caso por, costos, tiempo o recursos tecnológicos.

Asumiendo la complejidad y confiabilidad de conocer un método factible para detectar el engaño, Sullivan y Frank citado en el estudio cognitivo de la mentira humana en 2013, estipulan que existe evidencia que respalda el análisis del tono de voz para determinar la mentira de las personas, sobre esa base, los investigadores encontraron que las sonrisas y el tono de voz de quien decía la verdad o la mentira tenían un efecto positivo sobre quien engañaba, debido al grado de estrés al que se somete la fonación (Campos, Lopez, & Morales, 2013).

Considerando que la mentira es una herramienta que la utilizan muy a menudo las personas, existen campos profesionales como la criminología forense, la psicología, la medicina, la policía, las entidades gubernamentales, por nombrar las que mayor propósito requieren por conocer la verdad.

Justificación

En la actualidad son numerosos los estudios e investigaciones que buscan encontrar la metodología adecuada para determinar el engaño en las personas (Arias, 2018). Siendo el

campo forense el que ha enfatizado más en la efectividad de este procedimiento, muchos son los casos que han enfrentado agentes de criminología, o agentes policiales que se enfocan en resolver delitos como atracos, robo a vehículos, casas, entre otros.

Así mismo, actualmente el grupo del cuerpo de policía española acaba de implementar un nuevo miembro, es un algoritmo basado en inteligencia artificial llamada VeriPol, esta herramienta ha podido aumentar la efectividad de la resolución de casos que han sido falsos testimonios de gente que ha puesto denuncias por robos para cobrar sus seguros o tratar de encubrir que perdieron algún objeto personal, muchas veces los agentes dedicaban semanas a las investigaciones que en porcentajes muy bajos se descubrían que eran falsos, ahora esta herramienta que se basó en más de 1122 casos de delito por robo ha tenido una efectividad de un 25% más que el trabajo que realizaban los expertos dentro del departamento. Actualmente la policía española espera que esta nueva herramienta se implemente en sus demás comisarías, en las cuales se tiene una falta de personal considerable, se contempla entonces que las personas pensarán dos veces antes de tratar engañar a la policía. (Rodriguez , 2019)

Estos son algunos de los beneficios que ha causado el avance de la Inteligencia Artificial (IA) y se espera que esta herramienta se distribuya o se masifique a las diferentes áreas que prevalece en gran medida el tener certeza de la verdad que promulgan las personas.

De igual manera, en base a estudios preliminares sobre Machine Learning, en una investigación de Bravo Santiago denominada “Implementación de un sistema de reconocimiento automático de engaños mediante el análisis de la señal de la voz”, se realizó la predicción de engaños por medio del análisis de la señal de la voz a través de un sistema entrenado bajo aprendizaje supervisado mediante la utilización de la aplicación “Classification Learner” de la herramienta Matlab®, la cual permitió evaluar el porcentaje de exactitud del sistema

comparando diferentes tipos de algoritmos clasificadores tales como: “K-Nearest Neighbors” (KNN), “Support Vector Machine” (SVM), y Árboles de Decisión, de los cuales se obtuvieron como resultados de exactitud en hombres: 59.40%, 78%, y 71.90% respectivamente, así como 75%, 83%, 71.90% respectivamente, en mujeres. (Bravo, 2019)

Es por esta razón que se ha realizado esta investigación en base a Inteligencia Artificial usando Redes Neuronales Convolucionales de Deep Learning para contrastar con otros métodos de IA como Machine Learning y evidenciar el grado de confiabilidad, versatilidad y optimización de recursos, así de esta manera generar un precedente para que futuras investigaciones puedan determinar una mejor opción si pretenden enfocar sus algoritmos a ciertos campos, sea hacia la psicología, forense, policía, militancia, seguros, medicina, gobiernos, entre otros. Con la finalidad de siempre esclarecer la verdad frente a cualquier circunstancia y con un alto grado de confiabilidad (Rodriguez , 2019).

Objetivos

General

Implementar un algoritmo detector de mentiras/engaños mediante la utilización de Redes Neuronales Convolucionales como parte de Deep Learning.

Específicos

-Realizar un estudio previo de la utilización de las Redes Neuronales Convolucionales.

-Obtener una base de datos clasificada y etiquetada de verdades y mentiras e identificar las principales características de la señal de voz, mediante el estado del arte para la detección del engaño.

-Detectar o identificar características extralingüísticas (edad, sexo, entonación, micro temblores) de la voz y entrenar un sistema que use Redes Neuronales Convolucionales a través de la técnica de clasificación y prueba supervisada.

-Realizar técnicas de selección de características que otorguen los mejores resultados en la detección de engaños.

-Probar y evaluar el algoritmo aplicado bajo los parámetros de exactitud, precisión, sensibilidad y especificidad.

Capitulo II

Marco Teórico

Sistema Vocal Humano

La voz es una herramienta de comunicación entre personas, es el medio más básico para expresar y comunicar conocimientos, ideas y sentimientos. Se genera por la vibración del aire en la caja torácica a través de las cuerdas vocales ubicadas en la laringe, esto es un fenómeno fisiológico con propiedades acústicas. La voz percé, tiene diferentes propiedades acústicas entre las principales tenemos: la velocidad, el volumen, el ritmo, el tono, la duración y el timbre. Dichas características se las relaciona de una manera directa con la posición del cuerpo; influye de igual manera, el tono muscular y cómo se manejan algunas emociones (Nonó, Plaja, Pagés, Corbella, & Santamaria, 2015).

Al exhalar, el aire de los pulmones se fuerza a través de la glotis, lo que hace que vibren alrededor de 4 cuerdas vocales. La cavidad de la parte superior, está relacionada al sistema respiratorio y naso-faríngeo funcionando a modo de resonador (Sagasta, 2012).

El dispositivo que produce el sonido tiene la posibilidad de controlarse por la persona que genere un canto o que simplemente, habla. El cambio de cuán intenso es, está en manos de la cantidad de fuerza aplicada en la exhalación. En el hombre, las cuerdas vocales, tienen mayor longitud y grosor respecto a las femeninas, así como también a la de los niños, por lo que producen sonidos más graves (Sagasta, 2012).

El Sistema Vocal Humano (SVH) se compone del aparato respiratorio, de fonación y resonador que posteriormente se los analizará.

Aparatos del Sistema Vocal Humano

"La voz comienza con la inspiración (toma de aire), el aire llena los pulmones y luego sale a través de la espiración". (Son, 2010).

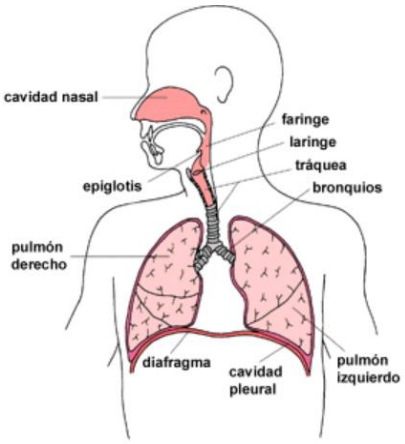
En el momento que se presenta una exhalación, el oxígeno ingresa a la tráquea desde los pulmones, luego ingresa a la laringe desde los pulmones y luego pasa por la mitad de los ligamentos de la laringe. A su vez dos musculaturas de tamaño reducido, se dilatan para obtener un tono alto y se relajan para obtener un tono bajo. De esta forma, la voz adquiere un tono. (Son, 2010).

Aparato Respiratorio.

De acuerdo a la Figura 1, se muestra que el aparato respiratorio está compuesto de varias partes (cavidad nasal, faringe, epiglotis, laringe, tráquea, bronquio, pulmón derecho, pulmón izquierdo, músculos intercostales, diafragma) que ayudan para que se produzca la voz (Gromé, 2019) .

Figura 1

Aparato Respiratorio



Nota. Tomado de la página Unprofesor, por (Gromé, 2019).

"Suministra el aire necesario para que se produzca la voz" (Gobierno de Canarias, 2016). La intención primordial del acto "respirar", es suministrar oxígeno a nuestro cuerpo, pero el mismo, a su vez, entra y sale, siendo también el motivo de producción del sonido (Son, 2010).

Aparato de Fonación.

Como se muestra en la Figura 2 dentro del aparato de fonación, es donde se produce puramente el sonido transformando al aire que se expulsa en la voz de la persona (Gobierno de Canarias, 2016).

Figura 2

Aparato de Fonación

Nota. Tomado de Pictoeduca, por (Guzmán, 2021).

Para producir las vibraciones sonoras, el oxígeno concentrado en los pulmones genera una sacudida, en primera instancia, esto se produce en la laringe; la cual está compuesta por una serie de cartílagos y ligamentos, acompañados de membranas que soportan los músculos de dicho tejido, a las cuales denominamos, cuerdas vocales (Ieda, 2016).

Para la doctora e investigadora Julia Máxima, en (2020) afirma que existen tres sistemas que hacen funcionar al aparato de fonación:

- Sistema de resonancia: Consta de tres cavidades articulares: faringe, cavidad bucal y cavidad nasal. El sonido que produce este sistema viaja desde las cuerdas vocales hasta las fosas nasales y la boca, provocando que el sonido se modifique y amplifique.

- Sistema de generación: este sistema se caracteriza por un flujo de aire excesivo a través de los músculos abdominales y del pecho, lo que aumenta la presión sobre los pulmones. Y de esta forma excitar el sistema de vibración.

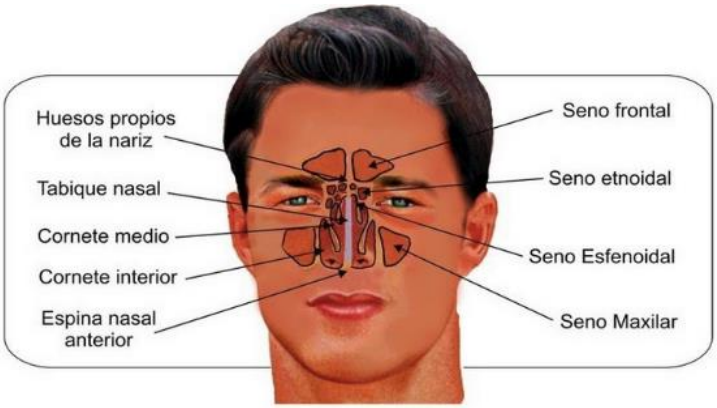
- Sistema de vibración: Está compuesto por cuerdas vocales clasificadas como superiores e inferiores, de las cuales solo la inferior está involucrada en la producción de sonido.

Aparato Resonador.

Conforme a la Figura 3, se define que: "El sonido que se genera en la laringe, a través de las cuerdas vocales, ayuda a conseguir un volumen más grande, debido a los resonadores naturales que posee el cuerpo humano" (Vozalia, 2020).

Figura 3

Aparato Resonador



Nota. Tomado de Universidad Camilo José Cela, por (Vozalia, 2020)

Son cavidades que asisten en la amplificación de los sonidos de manera natural. Conjuntamente, la anatomía personal de cada individuo (incluido el resonador) es lo que hace

que el sonido tenga un tono específico, por lo que logramos diferenciar las voces de varios individuos con diferente precisión, de modo que podamos saber quién está hablando sin verlos (Son, 2010).

El Engaño en SVH

Existen muchas fuentes de información en las que una persona puede basarse para detectar una mentira o engaño como: palabras, pausas, sonidos, expresiones faciales, movimientos de cabeza, gestos, posturas, respiración, rubor, complexión, sudor, etc.

Los mentirosos generalmente no pueden controlar todas sus acciones. Por el contrario, esconden y distorsionan, de esta manera, infieren que atraerá la mayor curiosidad de la gente: tienden a tener mucho cuidado al elegir las palabras. En la voz, integra todo lo que es el habla. Las señales más comunes en el sonido son pausas demasiado largas o pausas demasiado frecuentes o también el dudar cuando se empieza hablar (Ortiz, 2012).

Además, se debe considerar la entonación de la voz; cuando una persona está perturbada emocionalmente, el tono de la voz aumentará (más efectivo cuando el sentimiento es de ira o miedo) en cambio, la tristeza o el arrepentimiento bajan el tono. Según los experimentos realizados por, el tono aumenta con el engaño (Ekman, 2013).

De hecho, Puede ser más fácil ocultar los movimientos corporales que ocultar las expresiones faciales o los cambios de voz debidos a las emociones (Ortiz, 2012).

Sometimiento a Estrés.

Se suele identificar las respuestas al estrés como una preparación más o menos adecuada para afrontar situaciones de la vida (Bario, García, Ruiz, & Arce, 2016). ¨Cuando se presenta una dificultad, por cualquier razón, reaccionamos física y emocionalmente¨ (Centro Peruano de Audición, Lenguaje y Aprendizaje, 2020).

El sonido proporciona información de igual manera que un rostro, es decir, consigue revelar si un individuo es sensible, no obstante, aún no es conocida la posibilidad de entregar tanta información como la de un rostro que muestra la emoción exacta que se siente. (Ekman, 2013)

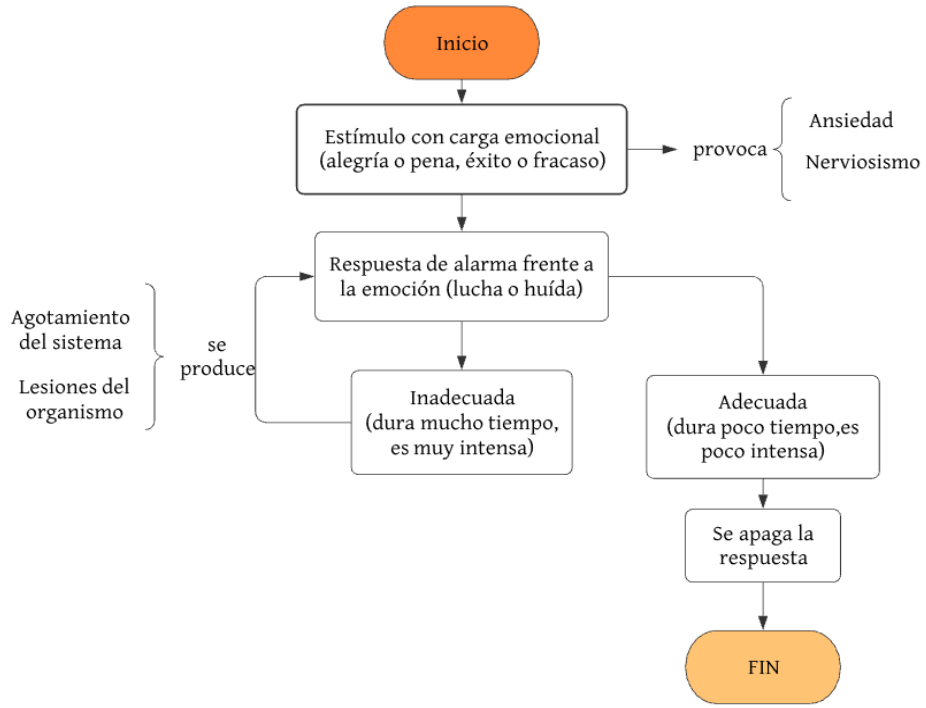
Los sonidos también están relacionados con áreas del cerebro involucradas en las emociones. Es difícil ocultar ciertos cambios en la voz que se producen al generar alguna alteración, así como también, la retroalimentación de la manera en que genera sonido una frecuencia de voz, lo cual es fundamental para un mentiroso.

Las personas siempre se sorprenden cuando escuchan su propia voz en una grabadora por primera vez, porque la parte de comprobación personal de la voz toma la ruta de dirección ósea, lo cual hace que suene distinto y sea increíble creer como el cuerpo, el organismo y todo el ser; actúa frente a la mentira o engaño.

A continuación, en la Figura 4, se presenta el proceso de cómo es la producción del estrés:

Figura 4

Flujograma de Producción de Estrés



Nota. Elaboración propia. Tomado de Redalyc, por (Bario, García, Ruiz, & Arce, 2016).

Detección del Engaño.

La respuesta más famosa al estrés es la respuesta de "lucha o huida" que se produce cuando se percibe una amenaza. Induce a que el cuerpo desencadena una liberación de algunas hormonas (entre ellas, cortisol y epinefrina, se la conoce también como adrenalina) hacia la sangre. Las hormonas mencionadas agrandan su capacidad de respuesta e intensidad. De igual manera pueden acelerar las palpitaciones cardíacas, elevar el índice de tensión arterial, fortalecer el régimen inmunológico y mejorar su memoria inmunitaria. Luego de plantarse a un estrés de plazo reducido, el cuerpo volvió a su estado normal (McEwen & Sapolsky, 2010).

Las condiciones de tensión pueden causar tensión general en la cabeza, el cuello, la garganta y el mentón, la zona del tórax superior, hombros, y espalda, y fluyen mediante la totalidad del aparato de sonido. (Llavina, 2013).

Los cambios de sonido provocados en la voz por las emociones no se ocultan fácilmente. Si lo que se quiere ocultar es la emoción que se siente en el momento de mentir, es probable que el mentiroso se auto revele.

Las detecciones más comunes del engaño dentro de lo vocal o la voz cuando una persona está mintiendo o está sometida a un grado de estrés son por lo general:

- El tono de voz: un estudio realizado a un grupo determinado de sujetos; arrojó resultados de que el 70% de ellos elevaba su tono de voz frente a una perturbación emocional (Ekman, 2013).

- Elección de las palabras: las personas que están al borde de la presión; analizan mucho las palabras que usarán para defenderse o responder frente al estímulo que les produce el estrés mismo (Ekman, 2013).

- En su defecto, mientras la presión aumenta abruptamente, las cuerdas vocales no funcionan eficazmente, lo cual produce cambios en la eficacia de la voz, especialmente en términos de tono, volumen y resonancia (Llavina, 2013).

“Con frecuencia lo que traiciona una mentira es la discrepancia entre el discurso verbal y lo que se pone de manifiesto en la voz, el rostro y el resto del cuerpo” (Ekman, 2013).

Estos resultados o detecciones comunes, se encuentran resumidas en los Indicios de Engaño y Reacción Frente a ello, mostrados en la Tabla 1 a continuación:

Tabla 1

Indicios de Engaño y Reacción Frente a Ello

Sospecha del Engaño	Información Descubierta
Deslices verbales	Pueden estar relacionados específicamente con una emoción; pueden delatar una información no relacionada con ninguna emoción.
Modo de hablar indirecto, circunloquios	Estrategia verbal no preparada de antemano, o bien presencia de emociones negativas, muy probablemente temor.
Pausas y errores en el habla	Estrategia verbal no preparada de antemano, o bien presencia de emociones negativas, muy probablemente temor.
Elevación del tono de voz	Emoción negativa, probablemente rabia y/o temor.
Disminución del tono de voz	Emoción negativa, probablemente tristeza.
Mayor volumen y velocidad del habla	Probablemente rabia, temor y/o excitación.
Menor volumen y velocidad del habla	Probablemente tristeza y/o aburrimiento.

Nota. Tomado de (Ekman, 2013) del libro ¿Cómo detectar mentiras?

En los cambios emocionales, fisiológicos o de sonido en el habla, se han creado varios aparatos que permiten su identificación y relacionarlos con el engaño. Estos dispositivos incluyen el “Evaluador de la Tensión Psicológica” (PSE), Analizador de Voz Mark II, Hagoth y Monitor de la tensión de voz. También el método Voice Stress Analysis (VSA) se basa en indicadores de bajo estrés, como microsismos o fluctuaciones de sonido, como indicadores indirectos de la mentira (Sanz, 2018).

Los aparatos mencionados otorgan la motivación necesaria de probar un algoritmo que basado en los mismos indicadores o características que ellos utilizan, mencionados en los capítulos subsiguientes, se podrá generar una solución de detección de engaños.

Análisis del Estrés en la Voz.

Así como el estrés afecta a la frecuencia cardíaca, la frecuencia respiratoria y la respuesta eléctrica de la piel, se dice que el efecto del estrés relacionado con el engaño en las cuerdas vocales se mide de tal manera que se lo puede observar (Chi Zaldívar, 2012).

El proceso de detección de engaño dentro del Análisis del Estrés en la Voz (VSA por sus siglas en inglés) es similar al que se usa en las pruebas de polígrafo; los examinadores de la VSA utilizan entrevistas previas a la prueba para preparar a los sujetos para el siguiente intento.

El propósito de este proceso es establecer la credibilidad del procedimiento de prueba, establecer una relación armónica con el sujeto, observar el comportamiento humano, generar problemas para la prueba y romper posibles obstáculos (Hopkins, Benincasa, Ratley, & Griego, 2015).

El software del programa VSA fue creado para identificar cambios en el habla o la voz y así detectar engaños a través del programa (CON-SIPA, 2018). En comparación con el uso de un polígrafo para verificar, se dice que el equipo VSA puede medir los efectos fisiológicos del engaño con mayor precisión, menos entrenamiento, menos invasivo, menos costo y un descubrimiento más directo (CON-SIPA, 2018).

Para Hopkins, Benincasa, Ratley, & Griego existen dos maneras o sistemas para que el VSA funcione e identifique el engaño: Sistema basado en energia de la señal de voz y sistema basado en frecuencia.

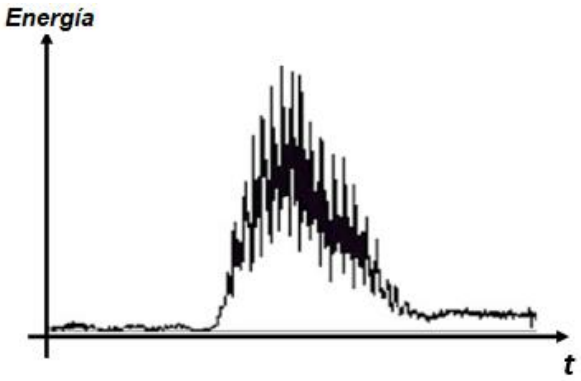
- ***Sistemas Basados en Energía de la señal de voz.***

Si una persona intenta hacer trampa, estará bajo presión. Este estrés dará como resultado una forma de onda medible que indica "trampa". Si la persona no tiene la intención de

hacer trampa, entonces no estará bajo presión y su respuesta producirá una forma de onda medible que indica "no hace trampa" (Chi Zaldívar, 2012). Dichas formas de onda, indican estrés bajo, medio y alto, como lo muestran la Figura 5, Figura 6 y Figura 7, respectivamente.

Figura 5

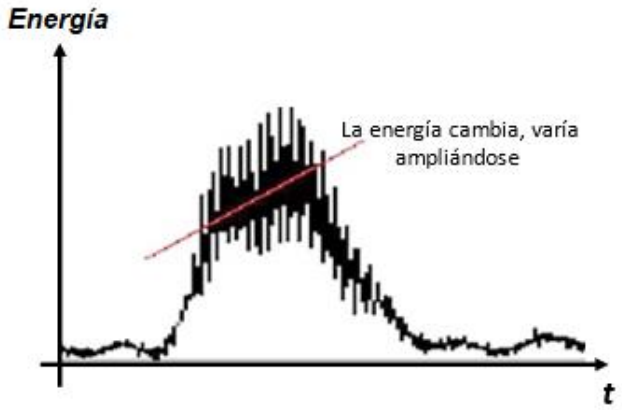
Forma de Onda Basada en el Sistema de Energía de la señal de voz que Indica: Bajo Estrés o Poco Estrés



Nota. Tomado de Proceedings of the 38th Hawaii International Conference on System Sciences, por (Hopkins, Benincasa, Ratley, & Griego, 2015).

Figura 6

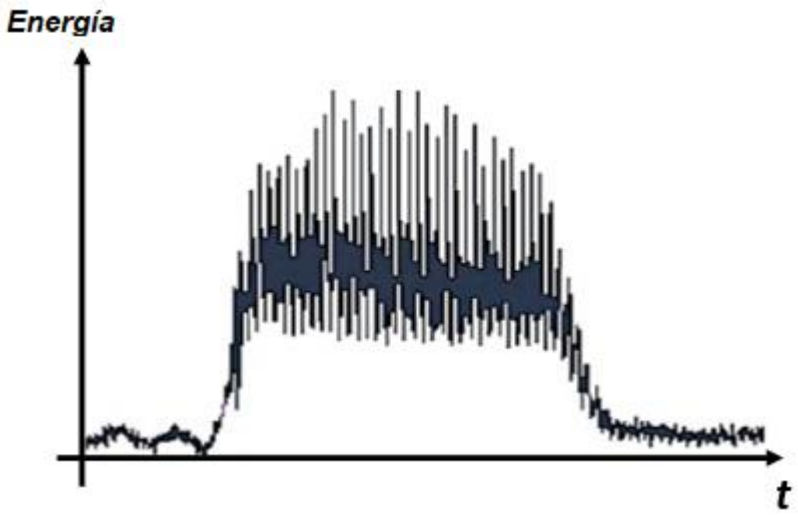
Forma de Onda Basada en el Sistema de Energía de la señal de voz que Indica Estrés Medio



Nota. Tomado de Proceedings of the 38th Hawaii International Conference on System Sciences, por (Hopkins, Benincasa, Ratley, & Griego, 2015).

Figura 7

Forma de Onda Basada en el Sistema de Energía de la señal de voz que Indica estrés Alto



Nota. Tomado de Proceedings of the 38th Hawaii International Conference on System Sciences, por (Hopkins, Benincasa, Ratley, & Griego, 2015).

De acuerdo a las figuras anteriormente mostradas, se observa que la forma de la energía cambia, ampliándose dependiendo de la cantidad de estrés identificada en la señal de voz, siendo así una característica favorable para la detección de engaños.

-Sistemas basados en Frecuencia

Las señales de voz reaccionan ante los cambios y por ello cabe destacar que, en estos tipos de sistemas, se puede determinar un rango de tensión continua y las posiciones de tensión relativas dentro de ese rango pueden ser comparadas entre sí, indicando en su resultado si la respuesta es o no engañosa.

Vibración de los Músculos.

El análisis del estrés de la voz se deriva del siguiente concepto: cuando una persona está bajo estrés, especialmente cuando el entrevistado está en peligro, el cuerpo aumentará la respuesta muscular para prepararse para la pelea y saltar (Xianfeng, 2005).

En 1971, Olof Lippold, un investigador del University College de Londres, publicó sus hallazgos en el Bulletin of Scientific American, titulado Physiological Tremors. Descubrió que las contracciones musculares voluntarias iban acompañadas de contracciones musculares involuntarias en el área de la garganta o laringe. Esto define al temblor como una contracción y relajación rítmica e involuntaria de dos paquetes musculares opuestos los cuales producen movimientos oscilantes de periodicidad habitual en una o más partes del cuerpo. Esta delimita una de las características primordiales de la Esclerosis Múltiple (EM) y forma parte de algunos de los síntomas diagnósticos de Charcot, acompañada de nistagmo y examen del habla (Liji, 2019).

Temblores de los Micro Músculos.

Lippold midió el efecto de la contracción muscular espontánea sobre el microtremor (pequeña vibración). En el campo de la detección del engaño, su trabajo se interpreta como que engañar voluntariamente puede provocar temblores mensurables. Estos leves temblores pueden afectar las cuerdas vocales, por lo que un examen cuidadoso del patrón de sonido puede revelar un engaño. De hecho, muchos desarrolladores de "detectores de mentiras" de presión de voz no afirman ser capaces de detectar mentiras.

En cambio, dicen que su dispositivo es capaz de detectar micro-vibraciones (o ciertas deformaciones), que pueden estar relacionadas con la presión de intentar encubrir o hacer trampa (Chi Zaldívar, 2012).

Para (Hopkins, Benincasa, Ratley, & Griego, 2015), el análisis del estrés de la voz se deriva del hecho de que cuando una persona está bajo estrés, el temblor micro muscular se produce en los músculos que forman el tracto vocal y se transmite cuando se habla.

Cuando una persona está bajo tensión, aumentará la vibración de los músculos, incluidas las cuerdas vocales. La frecuencia de la vibración está entre 8 y 12 Hz. Cuando una persona está mintiendo, la vibración aumentará más que la frecuencia normal, lo que genera estrés en el habla.

Estas vibraciones inaudibles se denominan micro temblores en el sonido o voz. También se consideran las características audibles de la voz humana, como la frecuencia fundamental y la micro vibración , y el propósito es detectar el engaño registrando los acentos humanos (Bravo, 2019).

Todas estas características mencionadas son las principales que nos van a permitir realizar la detección o diferenciación entre las verdades y los engaños.

Procesamiento de la Señal de la Voz y Extracción de Características.

El presente proyecto se basa en la utilización de la voz humana, la cual es una señal analógica, por lo que es necesario digitalizarla; cuando se habla de digitalización de la voz se hace referencia al procesamiento y conversión de la señal a impulsos que están representados por un código binario, es decir 0 y 1 (Salinas, 2020).

Es así que, la señal de voz, entendida también como la señal analógica de entrada (forma de onda continua) debe pasar por tres etapas para obtener como salida una señal digital (forma de onda discreta o flujo de bits); dichas etapas son: muestreo, cuantificación y codificación que al mismo tiempo se apoyan para ello en la transformada discreta de Fourier (Beltrán, 2003).

La detección de presión se realiza observando el cambio en el conjunto de características del habla neutra en comparación con el habla producida bajo presión. La cantidad de características utilizadas en la investigación debe analizarse cuidadosamente, ya que muchas de ellas pueden contener mucha redundancia y alto costo computacional. Por el contrario, es posible que una pequeña cantidad no tenga suficiente información para caracterizar el engaño (Bravo, 2019) .

De igual manera, se tiene una explicación a detalle a través de un diagrama de bloques del proceso mencionado en el Capítulo III, sección Descripción General del presente trabajo de titulación.

Transformada de Fourier:

Se utiliza para transferir la señal al dominio de la frecuencia para obtener información no delimitada en el dominio del tiempo. Las ecuaciones de la transformada de Fourier se pueden observar en las ecuaciones 1,2,3 y 4.

Ecuación de Síntesis (Transformada inversa de Fourier)

$$X(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{j\omega t}\,d\omega \qquad (1)$$

Ecuación de Análisis (Transformada integral de Fourier)

$$X(\omega) = \int_{-\infty}^{\infty} X(t)e^{-j\omega t}\,dt \qquad (2)$$

Nota. Tomado de la Universidad de Alcalá, por (Acevedo, 2010)

Las ecuaciones 1 y 2 son la base de la Transformada de Fourier para señales continuas, sin embargo, debido a que la investigación se centra en la utilización de señales digitales, serán de ayuda las ecuaciones 3 y 4, respectivamente.

Ecuaciones Para la Transformada de Fourier

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{j2\pi kn}{N}} ; k = 0,1,2, \dots, N-1 \quad (3)$$

$$x_n = \frac{1}{N} \sum_{n=0}^{N-1} X_k e^{\frac{j2\pi kn}{N}} ; n = 0,1,2, \dots, N-1 \quad (4)$$

Donde:

x_n *es la señal digital de la cual se quiere obtener la* $Transformada\ de\ Fourier$.

N: *número total de muestras a obtener de la* DFT $(Transformada\ de\ Fourier\ Discreta)$.

X_k: *es la señal discreta* $(digitalizada)$*de la DFT.*

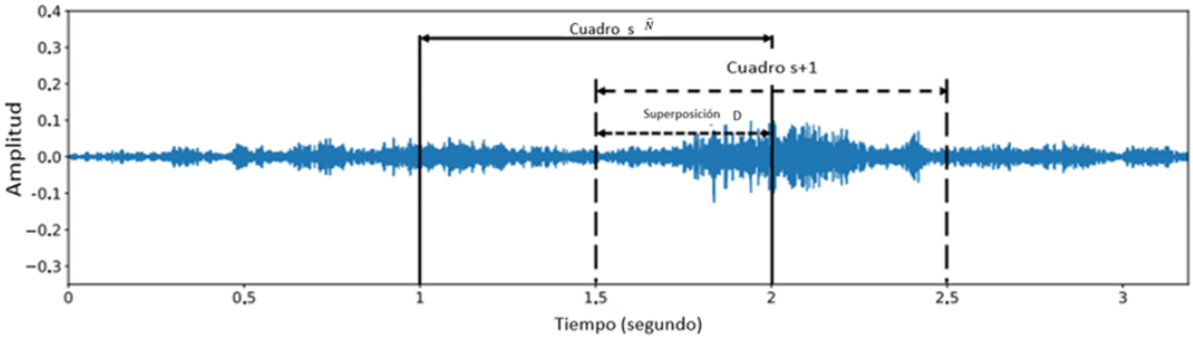
Nota. Tomado del Centro de Investigación en Matemáticas, por (Peña, 2019).

- <u>Audio Framming:</u>

Las señales de audio no son estacionarias, pero si usamos intervalos más pequeños en la señal, podemos esperar que sean fijos. Por tanto, calcularemos la transformada de Fourier en un pequeño intervalo de la señal original. El proceso de tomar estos intervalos se llama encuadre, y los intervalos son marcos. Por esta razón, el tamaño de ventana $\widehat{N}$ y el tamaño de deslizamiento D deben seleccionarse de manera que $D < \widehat{N}$, como lo muestra la Figura 8, de modo que la superposición de cada marco sea pequeña con la anterior. y el siguiente. Generalmente, para aplicaciones de voz, el tamaño de ventana $\widehat{N}$ está entre 20 y 30 milisegundos y el tamaño de superposición D, está entre 10 milisegundos (Peña, 2019).

Figura 8

Gráfico de explicación Audio Framming



Nota. Tomado de End-to-End Enviromental Sound Classification using a 1d Convolutional Neural Network, por (Koerich, 2019).

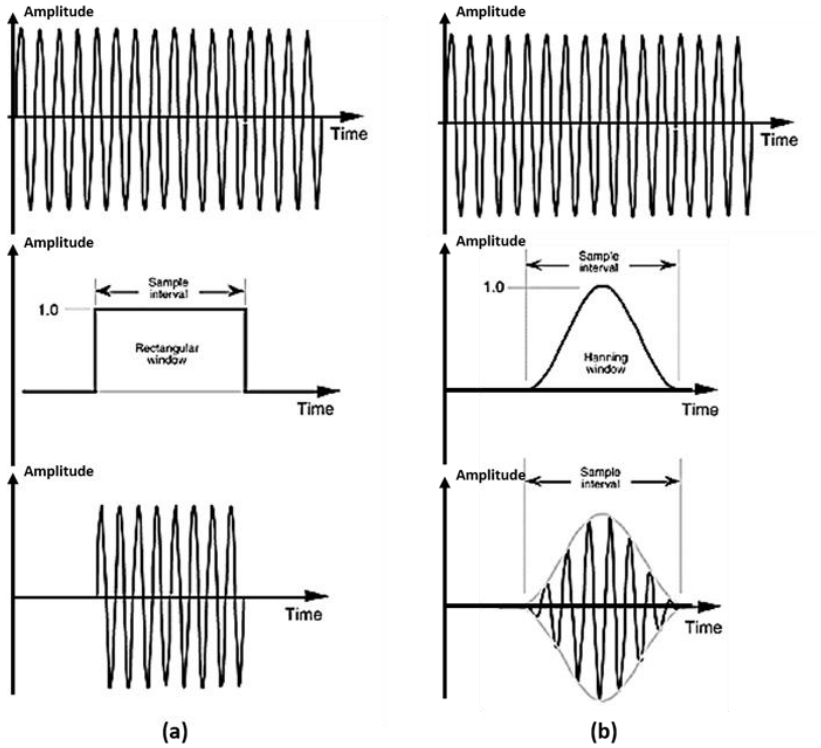
-<u>Eventanado:</u>

Debido a que las características de la voz son variantes, es necesario realizar el análisis de las mismas por pequeños “cuadros” o “intervalos”, el proceso de selección de estos intervalos se los conoce como “enventanado de la señal”, esto debido a que se desea eliminar el efecto de los bordes evitando “saltos” entre el final y el inicio de una ventana. (Peña, 2019)

En la Figura 9a se observa un ejemplo gráfico de enventanado mediante una ventana rectangular, mientras que en la Figura 9b se tiene el ejemplo de una ventana Hanning.

Figura 9

Ejemplos de Enventanamiento Rectangular y Hanning



Nota. Tomado de Proceedings of the 38th Hawaii International Conference on System Sciences, por (Lyons, 2004).

STFT

La STFT (Short-time Fourier transform) o en español mejor conocida como Transformada de Fourier de tiempo corto, es un método que permite la división de la señal en el dominio del tiempo en varias señales con duración menor para finalmente, cada señal obtenida, transformarla al dominio de la frecuencia (Ahmadizadeh, 2014).

Para ello la Transformada de Fourier de Tiempo Corto (STFT) es utilizada para estudiar las señales de voz debido a que son señales discretas. (Peña, 2019).

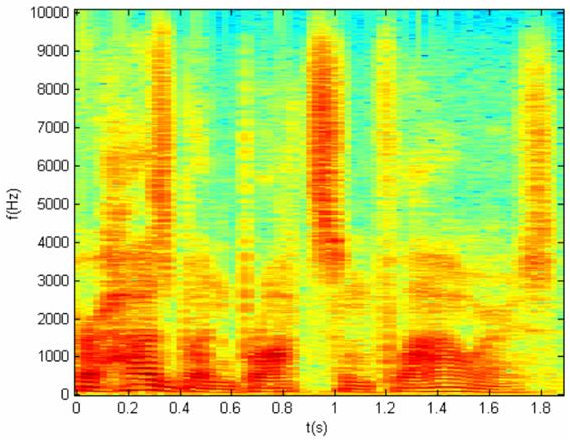
Espectrograma.

El espectrograma es una representación visual de las variaciones de la frecuencia de una señal identificada mediante los niveles de colores o grises que el sonido representa a lo largo del tiempo en el eje horizontal. Este espectrograma que surge a partir del resultado de la denominada "Transformada de Fourier de Término Corto" (STFT) aprovecha para examinar la duración, sonoridad, estructura del timbre, las pausas, la intensidad, y el ritmo. El proceso de la misma es la aplicación de la transformada de Fourier a una pequeña ventana de la señal y la combina con los resultados en una matriz bidimensional. Al utilizar una frecuencia adecuada de muestreo, una señal de audio puede seccionarse en varios fragmentos para transformarse independientemente. La matriz combinada entrega la relación tiempo-frecuencia, y los valores de cada cuadro muestra la magnitud de una frecuencia específica en un momento determinado. (Yang, 2018)

En la Figura 10 se evidencia un ejemplo de espectrograma de una señal de voz.

Figura 10

Ejemplo de Espectrograma – Señal de voz



Nota. Tomado de Proceedings of the 38th Hawaii International Conference on System Sciences, por (Martínez E. , 2006).

Este espectrograma representa de una forma visual las características mencionadas anteriormente para la diferenciación de señales de verdades y engaños.

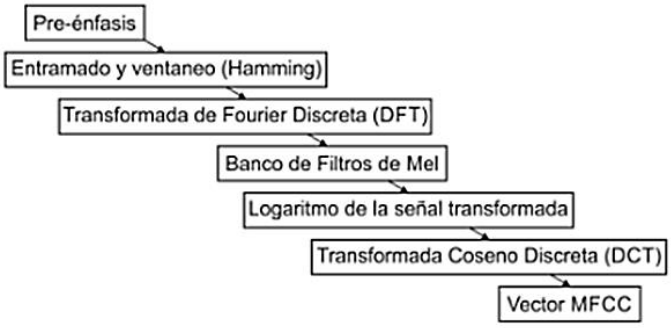
La construcción de los espectrogramas es producto de la desintegración del especto en segmentos discretos de tiempo de los valores de amplitud donde cada porción o ventana de tiempo es un conjunto de valores de densidad espectral en un intervalo de frecuencia (Araya, 2019).

Coeficientes de Mel.

Los Coeficientes Cepstrales en la Escala de Mel (MFCC) son la técnica más adecuada para el tratamiento y estudio de la voz debido a que simbolizan la amplitud del espectro de la voz de manera comprimida. A continuación, en la Figura 11, se demuestra el procedimiento de la elaboración de un vector característico de MFCC.

Figura 11

Proceso de obtención de vector característico de MFCC.



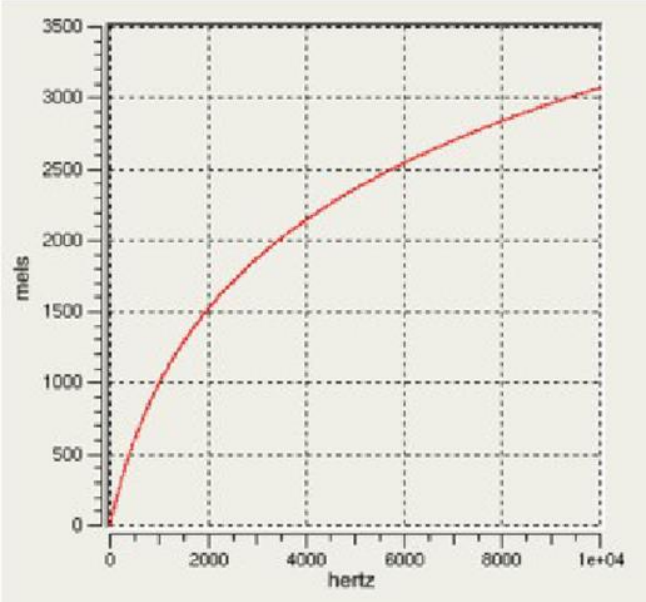
Nota. Tomado de Reconocimiento de Voz basado en MFCC, SBC y Espectrogramas, por (Martínez G. , 2013).

Una de las partes más importantes para diseñar cualquier sistema de reconocimiento de voz es la correcta extracción y selección de los parámetros de las señales de voz; esto afecta significativamente al rendimiento del reconocimiento. Una representación compacta sería proporcionada por un conjunto de Coeficientes de Cepstrum de Frecuencia Mel (MFCC), los mismos que son el resultado de la Transformada Coseno del logaritmo real del Espectro de Energía a Corto Plazo expresado en una escala de frecuencia Mel. Los MFCC han demostrado ser más eficientes en los sistemas de reconocimiento de voz. *(Tiwari, 2010)*

La peculiaridad característica que tienen es que en MFCC las bandas de frecuencia se encuentran ubicadas logarítmicamente, según la escala Mel, de manera que el punto de referencia se concreta equiparando un tono de 1000 Hz., 40 dBs por encima del umbral de audición del ser humano, con un tono de 1000 mels como lo muestra la Figura 12 *(Rincón, 2007).*

Figura 12

Gráfica de equiparación de un tono humano- Mel- Hertz.

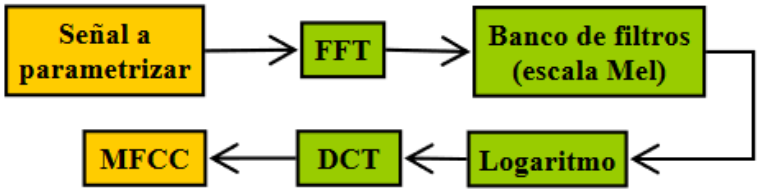


Nota. Tomado de Diseño, implementación y evaluación de técnicas de identificación de emociones a través de la voz, por (Rincón, 2007)

El procedimiento necesario para calcular los MFCC's se muestran en la Figura 13 detallada a continuación:

Figura 13

Diagrama de Bloques para el cálculo de MFCC's.



Nota. Tomado de Diseño de un sistema de reconocimiento del habla para controlar dispositivos eléctricos, por (Salcedo, 2006).

De este modo, en base al diagrama de bloques, se amplía el cálculo de los MFCC´s de la siguiente manera:

1) Se calcula la Transformada de Fourier de Tiempo Corto $X(n,\omega_k)$ a cada una de las tramas obtenidas de la etapa de preprocesamiento a través de la ecuación 5:

$$X(n,\omega_k) = \sum_{m=-\infty}^{\infty} x(m).w(n-m).e^{-j\omega_k m} \qquad (5)$$

$$Donde: \omega_k = \frac{2\pi}{N}.k$$

2) El cuadrado de la magnitud de $X(n,\omega_k)$ es ponderado por una serie de filtros distribuidos sobre la escala Mel para luego calcular la llamada "log-energía" del filtro l-ésimo por medio de la ecuación 6:

$$E_{Mel}(n,l) = \frac{1}{A_l}.\sum_{k=L_l}^{U_l} |V_L\,(\omega_K)\,.\,X(n,\omega_K)|^2 \qquad (6)$$

$Donde\ V_L\ representa\ el\ banco\ de\ filtros\ linealmente\ espaciado\ en\ la\ escala\ Mel$

Donde L_l y U_l son las frecuencias de corte inferior y superior del filtro l-ésimo.

3) Por último, el espectro logarítmico Mel es convertido reiteradamente al dominio del tiempo mediante la Transformada Discreta de Coseno (DCT), dado que los coeficientes cepstrales son números reales. De este modo, se demuestra el cálculo de estos coeficientes a continuación con la ecuación 7:

$$C_{Mel}[n,m] = \frac{1}{R} \cdot \sum_{l=1}^{R} log\{E_{Mel}(n,1)\} \cdot cos\left[n\left(l - \frac{1}{2}\right) \cdot \frac{\pi}{l}\right], \quad (7)$$

n = 1, 2, 3… K

Donde K es el número de coeficientes cepstrales, que por lo general se escoge entre 10 y 20. (Salcedo, 2006)

Métodos

Machine Learning.

Los costos operativos de alta complejidad tecnológica poseen mejoras y procesan una gran cantidad de datos. El aprendizaje automatizado generalmente se refiere a permutaciones en los procedimientos que generan labores relacionadas con la inteligencia artificial. Estas misiones envuelven reconocimiento, análisis, proyección, control, pronóstico y más (Bravo, 2019).

Así, Machine Learning permite la gestión de datos y optimiza la gestión de datos con el objetivo de reducir tiempo de actividad y costes. Con Machine Learning, puede encontrar patrones en sus datos con relativa facilidad, adaptar nuevos patrones si es necesario y aplicarlos a todos los conjuntos de datos (Bravo, 2019). Para poder comprender de mejor manera este sistema; se dará a conocer sobres los paradigmas del aprendizaje supervisado y no supervisado, y se los puede observar resumidos en la Figura 14.

Aprendizaje Supervisado.

Los sistemas de aprendizaje automáticos supervisados se originan en conocimientos previos (conjunto de adiestramiento), para por medio de esto poder pronosticar o determinar una etiqueta a casos nuevos. Es decir, el sistema de clasificación tomará experiencias previas para aprender, para así, después poder etiquetar la nueva información a analizar. Por tal motivo, en un sistema de clasificación se pueden diferenciar dos fases (Bravo, 2019).

A la primera fase se le conoce como de aprendizaje, y esta etapa se encarga de aprender las pautas o características de cada instancia. La segunda fase es conocida como test, esta es donde el clasificador es puesto a prueba haciendo que clasifique otro conjunto de instancias de las cuales ya se conoce la etiqueta correcta. De esta manera se podrá conocer la calidad o el comportamiento del sistema (Bravo, 2019).

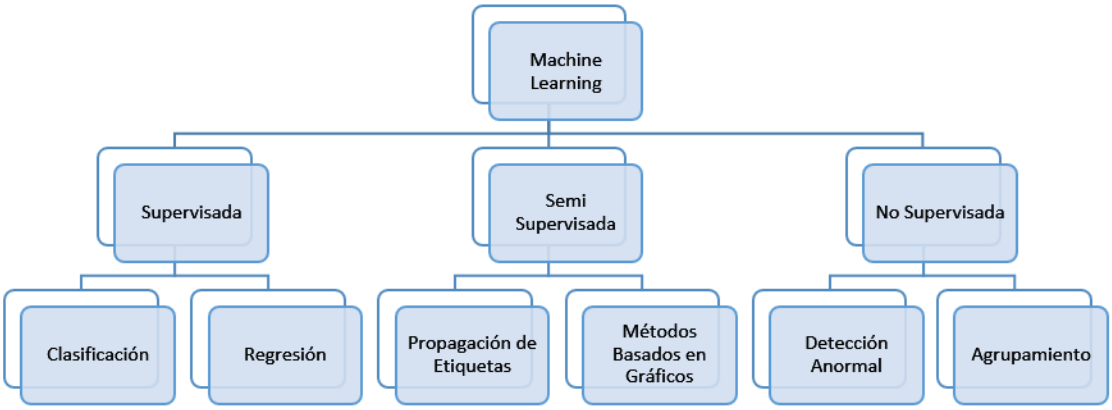
Aprendizaje No Supervisado.

El aprendizaje no supervisado o la agrupación en clústeres se basan en un grupo de instancias similares. El propósito de este proceso es extraer información o conocimiento relevante de los datos. Sin embargo, el concepto de parentesco tiene significados diferentes. Por tanto, existen diferentes técnicas de agrupamiento (Bravo, 2019).

El aprendizaje no supervisado se utiliza en áreas como la minería de datos, el aprendizaje automático, el reconocimiento de patrones y la biomedicina. Es por ello, que este método de aprendizaje ha tenido un aporte tan importante en los últimos años. El concepto de clúster intenta maximizar la distancia entre diferentes grupos de parentesco (Bravo, 2019).

Figura 14

Resumen de los Diferentes Paradigmas de Machine Learning



Nota. Tomado de la Universidad Politécnica de Madrid, por (Carreño, 2017).

Deep Learning.

El aprendizaje profundo es una nueva tecnología en el aprendizaje automático basada en la arquitectura de redes neuronales y es un tipo de Machine Learning. Está relacionado con algoritmos inspirados en la estructura y función del cerebro. Una red neuronal artificial se construye como un cerebro humano y los nodos de neuronas están conectados como una red.

El modelo de aprendizaje profundo se caracteriza por su estructura jerárquica, permitiendo métodos no lineales para procesar datos, pudiendo aprender a realizar tareas de clasificación directamente desde imágenes, texto o sonido (Itelligent, 2018).

Se utiliza una gran cantidad de datos etiquetados y una arquitectura de red neuronal con muchas capas para entrenar el modelo. La primera capa de la red neuronal procesa la entrada de datos sin procesar (como imágenes) y la pasa como salida a la siguiente capa. Este proceso se repetirá en secuencia hasta que se completen todas las capas de la red neuronal (Itelligent, 2018).

Un enfoque analítico tradicional es usar datos para seleccionar el diseño de las propiedades técnicas para seleccionar nuevas variables, luego calcular un modelo analítico y finalmente calcular los parámetros (o valores desconocidos) de este modelo. Estas técnicas pueden crear sistemas predictivos que generalmente se generan porque la integridad y la corrección dependen de la calidad del modelo y sus propiedades.

El nuevo enfoque para un aprendizaje profundo es reemplazar la formulación y especificación del modelo con caracterizaciones jerárquicas (o niveles) que aprende a reconocer las características diferidas de los ajustes de nivel.

Artificial Neural Networks

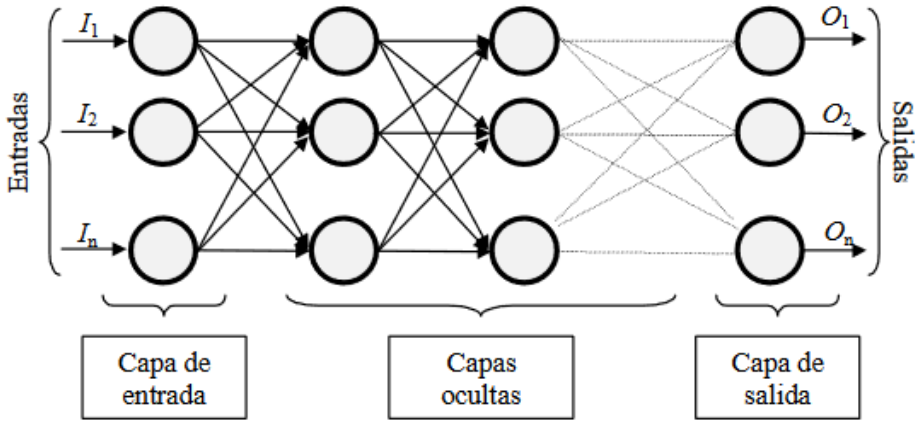
Las Redes Neuronales Artificiales son un modelo que intentan emular o reproducir el funcionamiento del cerebro humano y sus características propias, tales como: aprendizaje, memorización, etc. Para que exista una red neuronal es necesario un conjunto de unidades básicas conocidos como “nodos”, y a un nodo lo denominamos “neurona”, mismas que se encuentran conectadas y transmiten señales entre sí adquiriendo conocimiento a través de la experiencia.

Estas redes, se conforman por la interconexión de neuronas que se encuentran arregladas en tres capas o en otras ocasiones en más. Los datos o entradas $I_1, I_2, \dots I_n$ ingresan a través de la “capa de entrada”, pasan por la “capa oculta”, misma que puede estar compuesta de varias capas y finalmente se obtienen los valores de salida $O_1, O_2, \dots, O_n$ por la “capa de salida”. (Matich, 2001)

En la Figura 15, se observa la arquitectura de una Red Neuronal Artificial.

Figura 15

Arquitectura de una Red Neuronal Artificial



Nota. Tomado de Redes Neuronales: Conceptos Básicos y Aplicaciones, por (Matich, 2001).

Debido a que, en la sección de procesamiento de la señal de voz, se obtienen las características mencionadas como imágenes, se hace uso de las Redes Neuronales Convolucionales (CNN), ya que fueron diseñadas para funcionar con entradas de imágenes, motivo por el cual, dicho método es el más propicio de utilización.

Convolutional Neural Networks (CNN)

La era digital ha provocado la proliferación de diversas formas de datos de todo el mundo. Dichos datos no siempre están bien estructurados e incluso son inaccesibles. Una persona puede tardar algún tiempo en extraer manualmente información relevante de esta enorme cantidad de datos no estructurados. Sin embargo, existen algunas tecnologías de inteligencia artificial que pueden estructurar datos y extraer información útil de ellos. Por lo tanto, la empresa se da cuenta del enorme potencial que tiene la inteligencia artificial (Itelligent, 2018).

Las Redes Neuronales Convolucionales o Convolutional Neural Networks (CNN) son algoritmos que se utilizan en el aprendizaje automático para proporcionar funciones

informáticas de "visualización". Por eso, desde 1998, se ha podido clasificar imágenes, detectar automáticamente varios tipos de tumores, enseñar a los coches autónomos y muchas otras aplicaciones (Navia, 2018).

Las Redes Neuronales Convolucionales tienen mucha similitud a las redes neuronales ordinarias anteriormente mencionadas.

Las neuronas poseen, valores característicos, los cuales son receptores de una entrada con la que realizan un producto escalar y aplican una función de activación, además estas redes poseen una función de perdida en la última capa. No obstante, el objetivo principal de la utilización de las Redes Neuronales Convolucionales es el tratamiento de imágenes. A pesar que con las redes neuronales comunes se puede trabajar las imágenes, con CNN se genera una ganancia en cuanto a aumento de su tamaño, resolución y calidad. (Cortés, 2017)

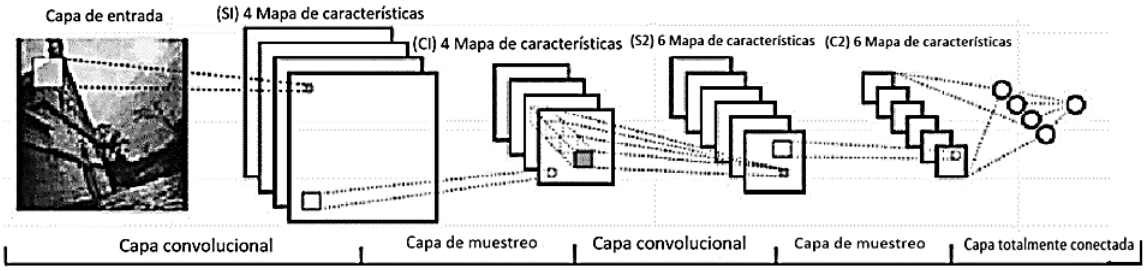
Para el entrenamiento de dichas redes pueden aplicarse los métodos de Machine Learning y Deep Learning antes mencionados, tales como, aprendizaje supervisado, y no supervisado.

Dichas redes toman este nombre debido a que su principal protagonista es la capa convolucional. Esta capa, a diferencia de las redes estándares que utilizan la multiplicación de matrices, realiza una operación denominada convolución. Misma que su funcionamiento se resume en recibir una imagen como entrada, aplicarle un filtro, el cual nos entrega un mapa de características reduciendo el tamaño de los parámetros (Cortés, 2017).

Dicho funcionamiento se puede observar en la Figura 16 .

Figura 16

Funcionamiento de la capa convolucional



Nota. Tomado de Herramientas Modernas en Redes Neuronales, por (Cortés, 2017)

Del funcionamiento se denota el significado de cada capa convolucional utilizada:

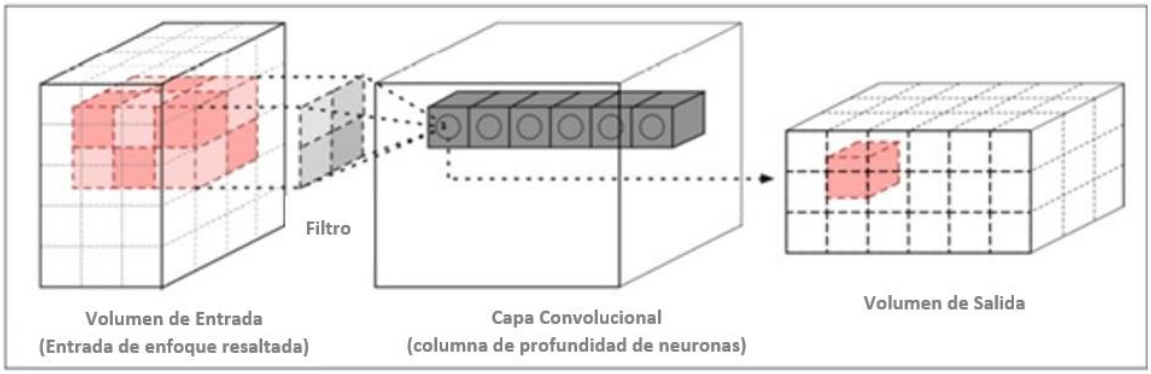
C: Convolution 2D Layer.

La capa convolucional calcula la convolución de su entrada bidimensional con un núcleo de tamaño fijo y luego la no linealidad de los elementos. La entrada puede contener múltiples canales del mismo tamaño, en este caso, usa un kernel separado para convolucionar cada canal y agregar los resultados. Del mismo modo, la salida puede constar de varios canales. Diferentes conjuntos de kernel. Comparado con la entrada, el kernel suele ser pequeño, lo que permite que la CNN maneje entradas grandes con pocos parámetros aprendibles. La capa convolucional transforma los datos de entrada mediante un conjunto de neuronas conectadas localmente desde la capa anterior. Esta capa calcula el producto escalar entre las regiones de neuronas en la capa de entrada y los pesos de sus conexiones locales en la capa de salida (Tayupanta, 2019).

En la Figura 17, se puede observar la entrada y salida de la capa convolucional mencionada.

Figura 17

Capa Convolucional con Volúmenes de Entrada y Salida



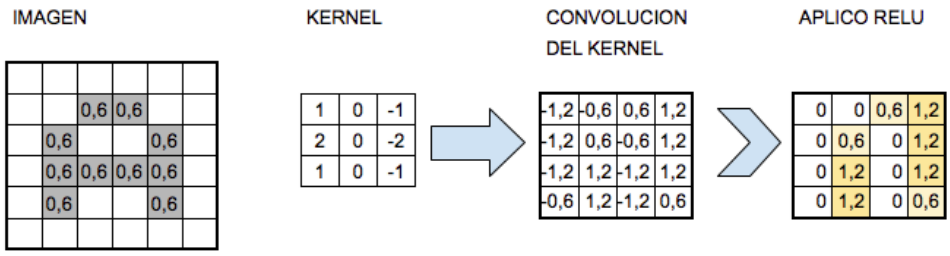
Nota. Tomado de Universidad de las Fuerzas Armadas "ESPE", por (Tayupanta, 2019).

R: Capa ReLU.

La capa ReLU realiza una operación de umbral en cada elemento de la entrada, y cualquier valor menor que cero se establece en cero (MathWorks, 2021). De acuerdo a la Figura 18, podemos observar el funcionamiento de la capa ReLU.

Figura 18

Funcionamiento de la Capa ReLU



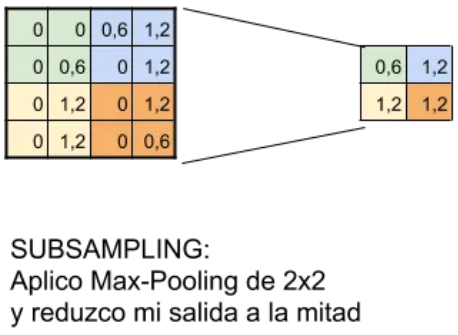
Nota. Tomado de "Aprende Machine Learning", por (Bagnato, 2011)

P: Max Pooling 2D Layer.

La capa de agrupamiento, mostrada en la Figura 19, generalmente se inserta entre capas convolucionales sucesivas para reducir gradualmente el tamaño del espacio (alto y ancho) representado por los datos en la red y ayudar a controlar el ajuste excesivo. La idea básica es dividir cada mapa de características para mosaicos del mismo tamaño, la capa de agrupación funciona de forma independiente en cada segmento de la entrada, tomando las submuestras de la capa convolucional y alimentándolas a la siguiente capa. Específicamente, se crea una celda para cada mosaico y el valor máximo en el mosaico se calcula y se difunde. Este gráfico de entidad máxima corresponde a la distribución de ponderación del valor en la celda y el esquema de agrupación permite a CNN generar propiedades de conservación como la invariancia de traducción (Tayupanta, 2019).

Figura 19

Funcionamiento de la Capa Max Pooling



Nota. Tomado de “Aprende Machine Learning”, por (Bagnato, 2011)

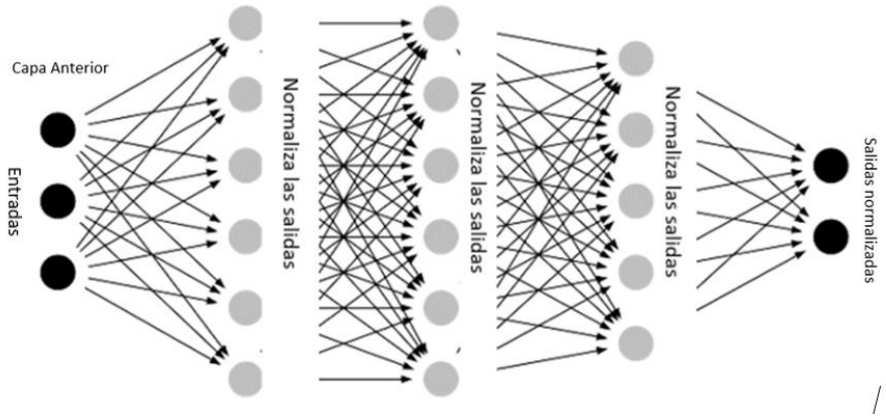
B: Batch Normalization Layer.

La normalización por lotes es una técnica utilizada para entrenar redes neuronales muy profundas, que normaliza la entrada de cada mini-lote a una capa. Esto tiene el efecto de estabilizar el proceso de aprendizaje y reducir en gran medida el número de ciclos de

entrenamiento necesarios para entrenar la red profunda (Brownlee , 2019). Una muestra del funcionamiento de la capa Batch, se evidencia en la Figura 20.

Figura 20

Muestra de la capa Batch Normalization Layer



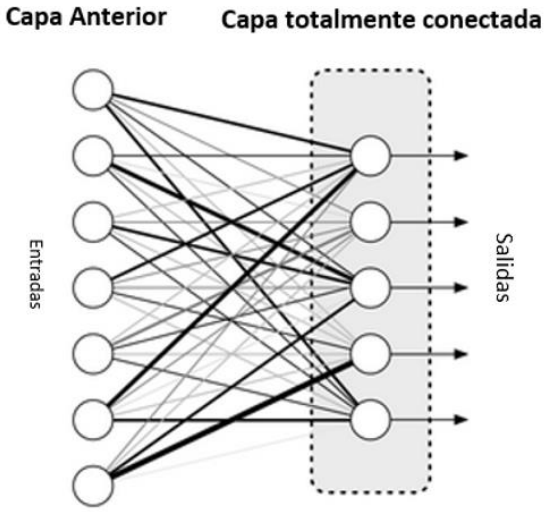
Nota. Tomado de “Aprende Machine Learning”, por (Bagnato, 2011)

F: Fully Connected Layer.

Finalmente, la capa completamente conectada, que se muestra en la Figura 21, descarta cualquier disposición espacial de su entrada transformándola en un vector, calcula el producto del punto y la matriz de ponderación y aplica la no linealidad del elemento al resultado. Por lo tanto, a diferencia de otros tipos de capas, no se limita a las operaciones locales y se puede utilizar como la etapa final de la integración de todas las capas. Información para la toma de decisiones (Tayupanta, 2019).

Figura 21

Muestra de la capa Fully Connected



Nota. Tomado de “Aprende Machine Learning”, por (Bagnato, 2011)

Capitulo III

Metodología de Investigación

Enfoque de Investigación

Dentro de esta investigación, se utilizó un enfoque de carácter mixto (cuanti-cualitativo), es decir que de manera cuantitativa se refleja en todo el capítulo IV al momento de aplicar el Software Matlab y recopilar datos numéricos que ayudan a determinar el uso de la inteligencia artificial y la detección de mentiras conforme a patrones numéricos (Sampieri, Fernández, & Baptista , 2010).

Método de Investigación

El método aplicado es de carácter deductivo ya que se parte de una red general de datos para llegar a datos más particulares y específicos donde se llegue a lograr crear el algoritmo para la detección de engaños (Sampieri, Fernández, & Baptista , 2010).

Diseño de Investigación

Se aplicó un diseño experimental. En este caso, se observa los resultados que el programa arroja y a partir de ellos se aplica una fórmula determinada para conseguir resultados donde aquellos no se alteran ni se interviene en el proceso para obtener aquellos datos (Sampieri, Fernández, & Baptista , 2010).

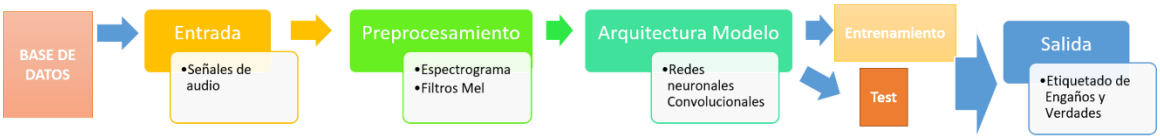
Descripción General

Este proyecto de investigación propone un algoritmo de etiquetado automático para detección de mentiras basado en contenido de señales de voz, que utiliza redes neuronales convolucionales utilizando la herramienta Matlab®.

La descripción general se puede encontrar en la Figura 22 y consta de 4 bloques, entrada, pre procesamiento, modelo de aprendizaje profundo y salida.

Figura 22

Diagrama de Bloques del Proyecto



Nota. Elaboración propia. Tomado de la Universidad de las Fuerzas Armadas "ESPE", por (Tayupanta, 2019).

Base de Datos

La base de datos utilizada es la que se presenta de manera pública, "RAVDESS", la cual contiene 7356 archivos de audio, etiquetados; dentro de esta base existen archivos que contienen señales con emociones, engaños y verdades, entre otras. Esta, fue basada en la que se encuentra alojada de manera pública bajo el nombre de "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)". La finalidad de la elaboración de la base de datos (RAVDESS), fueron estudios neuro científicos, psicológicos, psiquiátricos y ciencias de la computación, con el objetivo de detección de trastornos neurológicos (Livingstone & Russo, 2018).

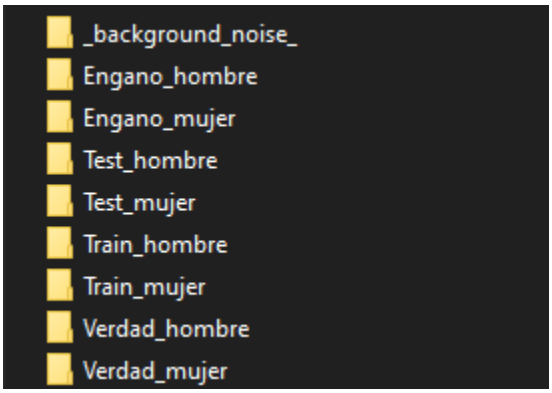
Del total de audios de la base, se utiliza dentro del trabajo de titulación, un total de 192 participantes. La base de datos tiene un total de 96 audios grabados por hombres y 96 audios grabados por mujeres, debidamente etiquetada para el estudio a realizarse, los cuales fueron tomados de (Bravo, 2019).

Con estos 192 archivos mencionados, tenemos nuestra base de datos con la que se va a trabajar en el presente proyecto.

La base de datos mencionada anteriormente está compuesta de 9 carpetas totales, de las cuales, 4 están clasificadas para Verdades y engaños, tanto para hombres como para mujeres, 2 para test de hombres y mujeres, 2 de entrenamiento o "*training*" de hombres y mujeres y 1 de ruidos adicionales, como lo muestra la Figura 23.

Figura 23

Muestra de la clasificación de carpetas de la Base de Datos



Nota. Elaboración propia

Los audios de la base presentan las siguientes características:

- Velocidad de bits 768 kbps
- Frecuencia 16KHz
- Duración 3 segundos

Cada una de las 4 carpetas de la clasificación de verdades y engaños consta de:

- Verdad Hombre: 48 audios grabados por hombres.
- Verdad Mujer: 48 audios grabados por mujeres.
- Engaño Hombre: 48 audios grabados por hombres.
- Engaño Mujer: 48 audios grabados por mujeres.

De esta manera se obtienen los 96 audios grabados por hombres y 96 por mujeres.

Cada uno de los 48 archivos de audio por carpeta, se encuentran numerados del 1 al 48 seguida de la letra “e” si corresponde a engaño, o la letra “v” si corresponde a una verdad. Un ejemplo de esta nomenclatura y numeración se pueden observar claramente en la Figura 24 y Figura 25, respectivamente.

Figura 24

Muestra de los archivos de audio contenidos en la carpeta “Engano_hombre” de la Base de Datos

Nota. Elaboración propia

Figura 25

Muestra de los archivos de audio contenidos en la carpeta “Verdad_mujer” de la Base de Datos

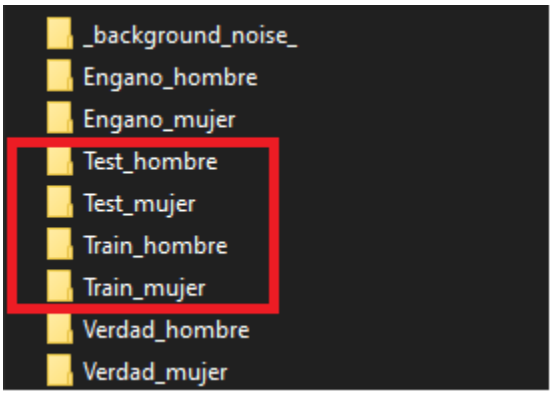
Nota. Elaboración propia

Base de Datos de Entrenamiento y Evaluación del Clasificador.

En este punto es necesario denotar las 2 carpetas de Test y 2 carpetas de Train para hombres y mujeres, como lo muestra la Figura 26.

Figura 26

Muestra de las carpetas de Test y Train de la Base de Datos



Nota. Elaboración propia

Para ejecutar el entrenamiento se incluye un conjunto de señales que pueden ser consideradas como positivas, las cuales se agregaron de la base de datos RAVDESS, para este proceso se tiene que, de la base de datos de 96 audios, 64 audios sirven para el entrenamiento y 32 para la evaluación, los cuales representan el 70% y 30% respectivamente del total de la base de datos, teniendo así 70% para *training* y 30% para *testing* por género. Para ello, cada carpeta de la clasificación de Test y Train consta de:

- Test_Hombre: 32 audios, los cuales corresponden a las numeraciones del 33 al 48 de los archivos de verdades hombre y del 33 al 48 de los archivos de engaños hombres.

- Test_Mujer: 32 audios, los cuales corresponden a las numeraciones del 33 al 48 de los archivos de verdades mujeres y del 33 al 48 de los archivos de engaños mujeres.
- Train_Hombre: 64 audios, los cuales corresponden a las numeraciones del 1 al 32 de los archivos de verdades hombre y del 33 al 48 de los archivos de engaños hombres.
- Train_Mujer: 64 audios, los cuales corresponden a las numeraciones del 1 al 32 de los archivos de verdades mujeres y del 33 al 48 de los archivos de engaños mujeres.

Un ejemplo de esta clasificación y numeración de los archivos se pueden entender de mejor manera en la Figura 27 y Figura 28, respectivamente.

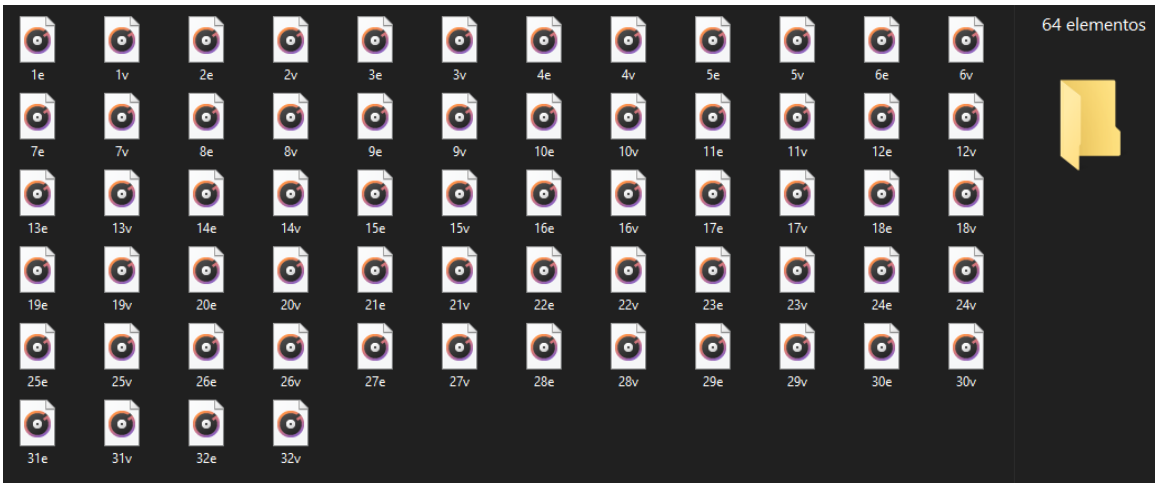
Figura 27

Muestra de los archivos de audio contenidos en la carpeta "Test_Hombre" de la Base de Datos

Nota. Elaboración propia

Figura 28

Muestra de los archivos de audio contenidos en la carpeta "Train_Mujer" de la Base de Datos



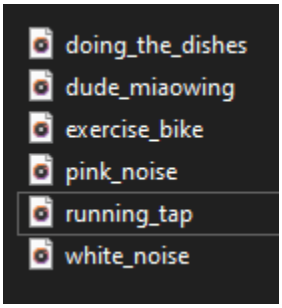
Nota. Elaboración propia

Finalmente, la carpeta de ruidos, denominada "background_noise", consta de 6 archivos de audio de ruidos externos, como: lavar platos, ejercicio en bicicleta, maullidos de gatos, etc, que nos servirán para evaluar de mejor manera el desempeño del clasificador, debido a que la red debe poder no solo reconocer las diferentes señales de voz, sino también detectarlas si a las señales de entrada agregamos diferentes ruidos de fondo, para ello, el algoritmo genera 100 señales de background que se analizan en conjunto con el entrenamiento y se reconoce la cantidad de las mismas que fueron detectadas en las señales de la base de datos de hombres y mujeres.

Estos archivos se pueden observar en la Figura 29.

Figura 29

Muestra de los archivos de audio contenidos en la carpeta "background_noise" de la Base de Datos



Nota. Elaboración propia

Entrada

La entrada está conformada por las señales de audio que provienen de la base de datos descrita y detallada en el inciso anterior, dichas señales de audio son las que ingresarán al sistema a la etapa de pre procesamiento, debido a que, para mejorar el rendimiento del sistema, es muy importante contar con un pre procesamiento adecuado de los datos de entrada.

Pre Procesamiento de las Señales de entrada.

El pre procesamiento que se realiza es la extracción del espectrograma de la señal de entrada, esto va a producir una imagen (ver Figura 31), la cual permite representar digitalmente la información contenida en la señal de audio. Este procesamiento se ejecuta a toda la base de datos debido a que nuestra CNN necesita como entradas las imágenes del espectrograma obtenidas en el presente pre procesamiento. Un esquema de este proceso, se muestra en la Figura 30.

Figura 30

Esquema General del Procesamiento

Nota. Elaboración propia.

Para dicho proceso, lo que se realiza primero es dividir la señal de audio en tramas y una función de ventaneo el cual en este caso es una ventana de Hamming de 25ms, lo que permite esta ventana es eliminar los bordes de la señal y resaltar el centro de la trama.

Dentro del procesamiento de la señal, en primera instancia se toman los audios de la base de datos, creando un array con dichas señales de audio que serán convertidas en imágenes mediante la utilización de espectrogramas Mel, teniendo así la representación digital de la señal de audio, para aplicar la Transformada Rápida de Fourier (STFT) obteniendo la amplitud del espectro con el objetivo de pasarlo al dominio Mel a través de un banco de filtros.

Para mejorar estos datos y obtener una distribución más uniforme se utiliza un desplazamiento llamado epsil de 1e-15. En la Figura 31 y Figura 32 se pueden observar los gráficos obtenidos de este pre procesamiento en una señal de ejemplo de hombres.

Figura 31

Señal de Voz y Espectrograma Arrojados por Matlab del Entrenamiento (Hombres)-Verdades

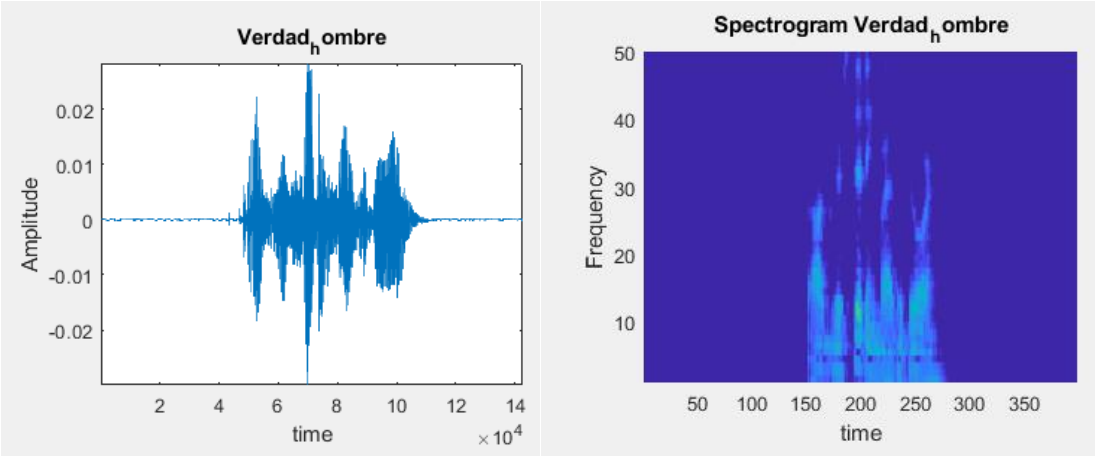
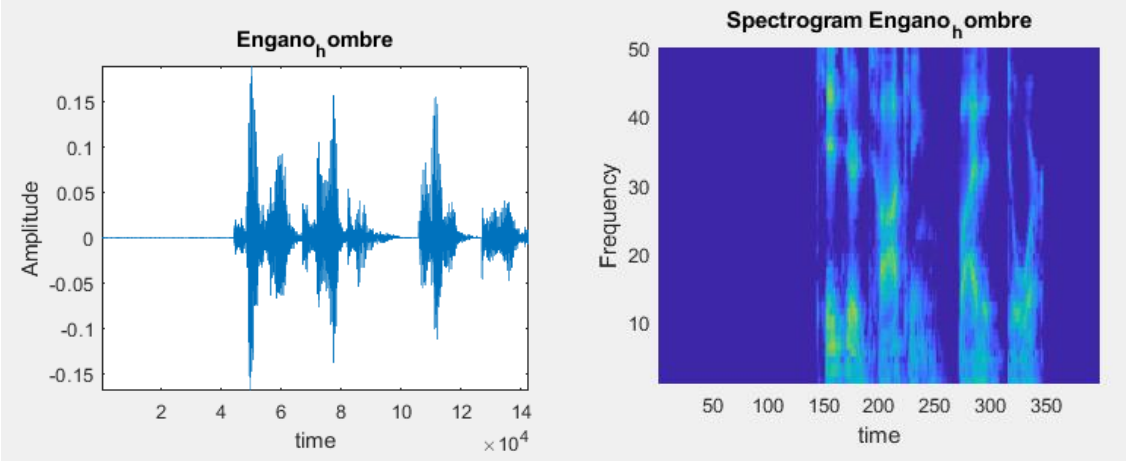


Figura 32

Señal de Voz y Espectrograma Arrojados por Matlab del Entrenamiento (Hombres)-Engaños



De la gráfica mostrada se observa, la amplitud de la señal de voz ingresada al algoritmo, y el respectivo espectrograma de Mel en hombres. De igual manera se presentan a continuación en las Figuras 33 y 34 las gráficas obtenidas en una señal de ejemplo para mujeres.

Figura 33

Señal de Voz y Espectrograma Arrojados por Matlab del Entrenamiento (Mujeres)-Verdades

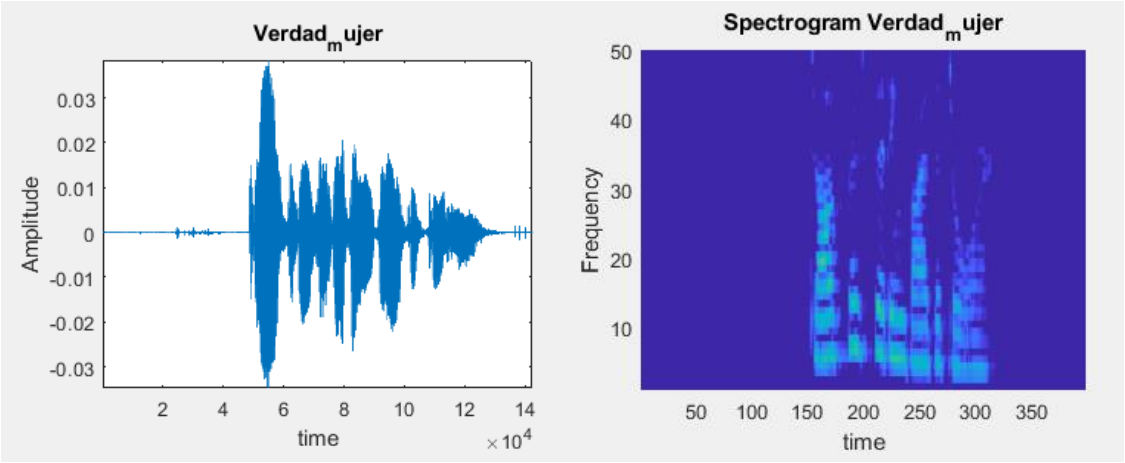
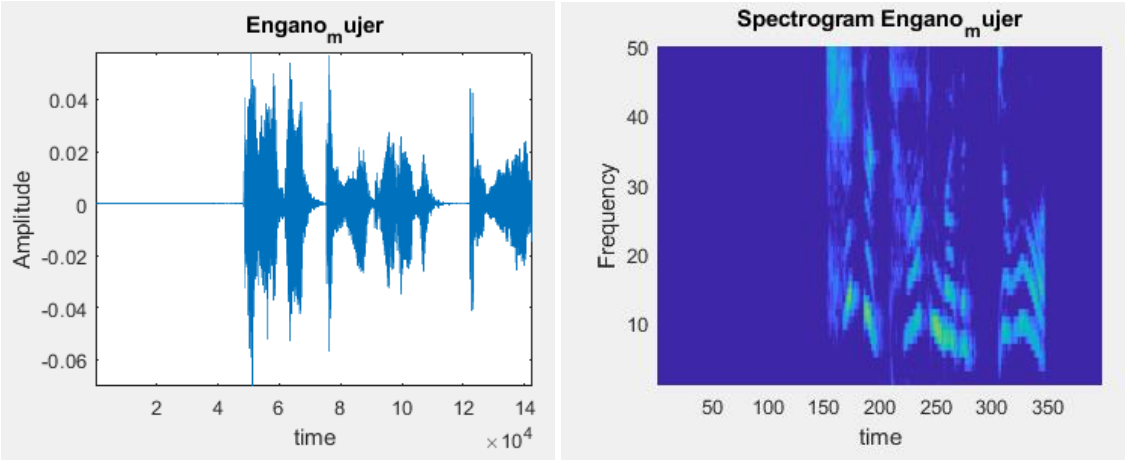


Figura 34

Señal de Voz y Espectrograma Arrojados por Matlab del Entrenamiento (Mujeres)-Engaños



Arquitectura Propuesta.

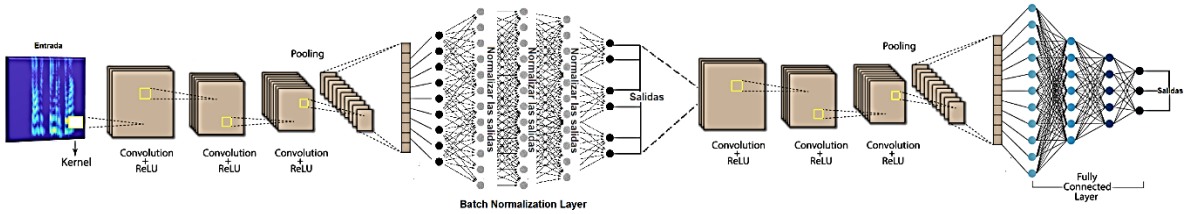
Para la arquitectura del presente trabajo se pretende crear una Red Neuronal Convolucional (CNN) que nos permitirá identificar los engaños, la cual tendrá como entrada las imágenes obtenidas del espectrograma del pre procesamiento mencionado en el inciso anterior y como salida obtendremos los resultados de detección de engaños, para ello, se propone obtener resultados de 2 “modelos”, un modelo se refiere a la aplicación de diferentes

cantidades de capas convolucionales de la CNN, teniendo así el primer modelo de 5 capas que se puede observar en la Figura 35 y el segundo modelo de 10 capas convolucionales en la Figura 36; ambos modelos fueron utilizados con la finalidad de obtener un contraste que nos ayude en la visualización de como un modelo de mayor cantidad de capas conseguiría mejores resultados en la realización de los presentes experimentos, utilizando un orden de capas convolucionales tales como:

- **Modelo 1:** 5 capas: CR-CR-CRPB-CR-CR- CRPF.
- **Modelo 2:** 10 capas: CR-CR-CRPB-CR-CR-CRPB-CR-CR-CRPB-CRPF

Figura 35

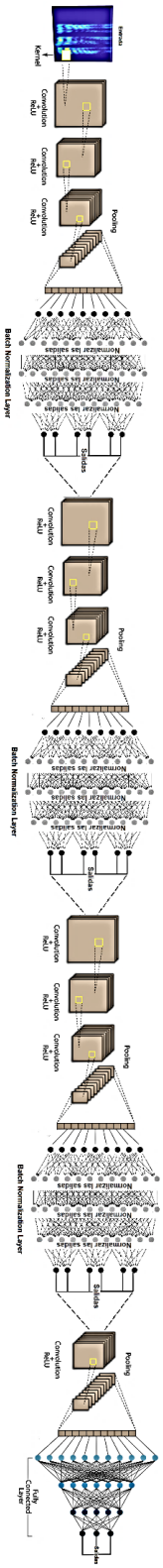
Arquitectura de la CNN del Modelo 1 – 5 capas



Nota. Elaboración propia

Figura 36

Arquitectura de la CNN del Modelo 2 – 10 capas



La CNN, toma la imagen obtenida en el preprocesamiento (espectrograma), la divide en cuadrículas, y la va operando matemáticamente (producto escalar) de izquierda a derecha y de arriba abajo, con una pequeña matriz de valores aleatorios denominada “kernel”, generando una nueva matriz que se conoce como convolución del kernel, a la misma aplicamos la capa ReLU, la cual toma una matriz solamente con valores positivos, y dichas salidas generan el llamado “mapa de características”, que nos servirá para el reconocimiento, los mapas vuelven a generar el proceso de convoluciones, pasando estas salidas a la capa de Max Pooling quedándose con los valores más altos de dichas matrices, para alimentar la capa de Batch Normalization Layer obteniendo dichas características normalizadas que alimentarán un nuevo proceso de convolución, ReLu, Max Pooling, y dichas salidas alimentan la capa totalmente conectada para así conseguir finalmente nuestros resultados de detección.

Entrenamiento y Evaluación de la Red.

Para continuar con el proyecto, las redes neuronales descritas en la sección anterior deben ser entrenadas, para lo cual, utilizamos la base de datos conocida precedentemente, para obtener ambos modelos entrenados en 25 épocas, de donde una época describe la cantidad de veces que un algoritmo observa el conjunto de datos en su totalidad.

Para ello, se utilizó la optimización adaptativa de Adam, el cual es un algoritmo de optimización que se utiliza en lugar del procedimiento de descenso de gradiente estocástico clásico para actualizar los pesos de red de forma iterativa en función de los datos de entrenamiento (Kingma & Ba, 2015).

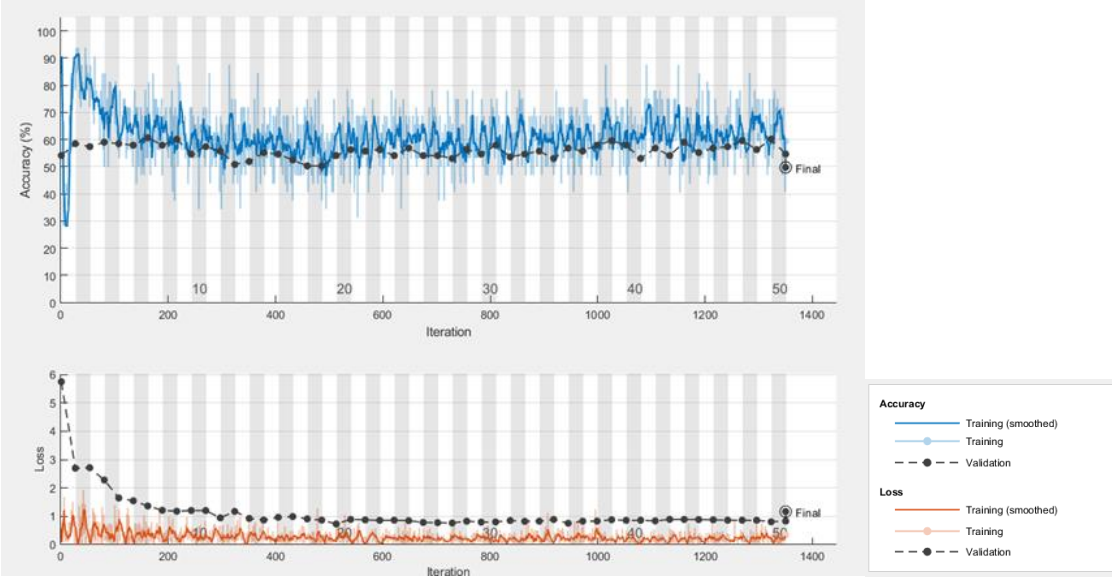
En los experimentos, a su vez, se realiza la variación del *mini-batchSize* a tamaño 32 y 128, adicional a esto se detalla que se realizó el entrenamiento durante 25 épocas, esto debido al valor recomendado por el ejemplo de entrenamiento de Matlab el cual menciona que al pasar

20 épocas se reduce la tasa de aprendizaje por un factor de 10 (MathWorks, 2021), por lo tanto, éstos datos y valores permiten el entrenamiento óptimo final de la red con los datos de training y validation con el valor de batchSize máximo utilizado de 128.

Dicho proceso de entrenamiento se lo aplica para el modelo 1 y 2, respectivamente, obteniendo así ambos modelos entrenados, para ello Matlab, nos entrega la gráfica respectiva de la precisión del entrenamiento y pérdidas como lo muestra la Figura 37.

Figura 37

Ejemplo de Entrenamiento Red Neuronal (Hombres)



De la Figura mostrada, se observa el porcentaje de precisión que obtiene el entrenamiento que se encuentra denotado en línea continua de color azul, respecto a los datos de validación que se muestran en línea punteada de color negra. Ambas líneas se encuentran muy cercanas, por lo que gracias a ello se denota si la precisión del entrenamiento obtuvo un buen resultado cuando llegó a su punto mostrado como “Final”. De igual manera se evidencia una gráfica con las pérdidas, que, al igual que en la precisión, podemos comparar su variación respecto a la validación en línea punteada.

Como resultado del entrenamiento, y en vista de que empleamos un algoritmo, es necesario la utilización de una “matriz de confusión”, la cual es una herramienta que nos ayuda a la visualización del desempeño de un algoritmo empleado en aprendizaje supervisado, la misma se observa en la Figura 38.

Figura 38

Parámetros de una Matriz de confusión

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

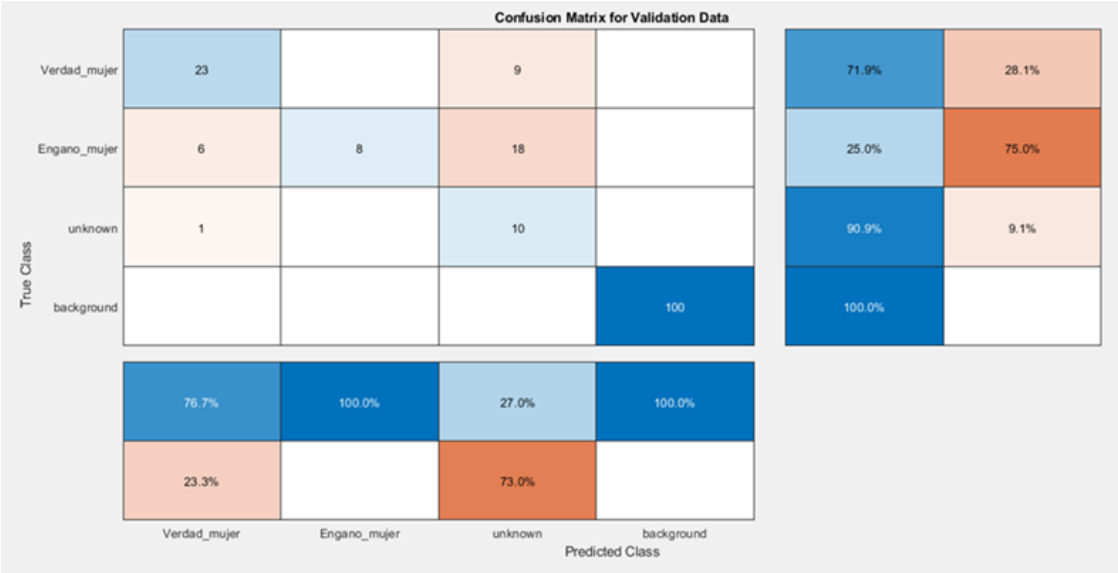
Nota. Elaboración propia

Por tal motivo, adicional a las gráficas de entrenamiento, como se puede observar en la Figura 39, Matlab nos entrega a su vez las respectivas matrices de confusión de dichos experimentos para la obtención de los parámetros mencionados a continuación:

- **Verdaderos Positivos:** Un audio es ingresado como verdad y el algoritmo lo clasifica como verdad. (Dependiendo en cada clase, siendo Verdad_hombre o Verdad_mujer)
- **Falsos Positivos:** Un audio es ingresado como engaño y el algoritmo lo clasifica como verdad. (Dependiendo en cada clase, siendo Verdad_hombre o Verdad_mujer)
- **Verdaderos Negativos:** Un audio es ingresado como engaño y el algoritmo lo clasifica como engaño. (Dependiendo en cada clase, siendo Engaño_hombre o Engaño_mujer)
- **Falsos Negativos:** Un audio es ingresado como verdad y el algoritmo lo clasifica como engaño. (Dependiendo en cada clase, siendo Engaño_hombre o Engaño_mujer)

Figura 39

Ejemplo de Resultado de Matriz de Confusión (Mujeres)



De la Figura 39 mostrada se denota que las clases evaluadas son las carpetas de nuestra base de datos, tanto para hombre como para mujer, en el presente experimento se observa: Verdad Mujer, Engaño Mujer, desconocidas (cuando no detecta si fueron verdades o engaños), y los ruidos externos agregados de la base de datos mencionada con anterioridad.

Se denota en nuestra matriz, la clase “background” simplemente por tener un control de cuántas de las 100 señales de background noise explicadas en la sección Base de Datos, se tomaron o fueron detectadas en las clases evaluadas de verdades y engaños.

Para aclarar la obtención y significado de los parámetros y valores que contiene la matriz de confusión, se muestra a continuación un detalle de la misma.

Obtención de Verdaderos Positivos

Dentro de nuestra matriz de confusión, los valores que se encuentran en la diagonal, representan nuestros verdaderos positivos para cada clase, es decir, VP0 representa el

Verdadero Positivo (VP) de la Clase 0, VP1 sería el verdadero positivo de la Clase 1 y VP2 el de la Clase 2, respectivamente. Estos valores se pueden observar con mayor facilidad en la Figura 40.

Figura 40

Obtención de Verdaderos Positivos de una Matriz de confusión de 3 clases.

Obtención de Verdaderos Positivos

Clase 0	**VP0**		
Clase 1		**VP1**	
Clase 2			**VP2**
	Clase 0	**Clase 1**	**Clase 2**

VP Clase 0=	VP0
VP Clase 1=	VP1
VP Clase 2=	VP2

Obtención de Verdaderos Negativos

Para la obtención de los Verdaderos Negativos (VN), es necesario centrarse en cada clase, para ello, si se desea el VN de la Clase 0, se debe excluir los valores situados en la fila y columna donde se encuentra el VP de dicha clase, y se realiza la suma de los valores restantes. Esta explicación se muestra con mayor facilidad en la Figura 41.

Figura 41

Obtención de Verdaderos Negativos de la clase 0 de la Matriz de confusión de 3 clases.

Obtención de Verdaderos Negativos Clase 0

Clase 0	**VP**		
Clase 1		**VN1**	**VN2**
Clase 2		**VN3**	**VN4**
	Clase 0	**Clase 1**	**Clase 2**

VN Clase 0=	VN1+VN2+VN3+VN4

De manera análoga, para obtener el VN de la Clase 1, se realiza el mismo procedimiento antes mencionado, y así, sucesivamente con todas las clases, como lo muestran las Figuras 42 y 43.

Figura 42

Obtención de Verdaderos Negativos de la clase 1 de la Matriz de confusión de 3 clases.

Obtención de Verdaderos Negativos Clase 1

Clase 0	VN1		VN2
Clase 1		VP1	
Clase 2	VN3		VN4
	Clase 0	Clase 1	Clase 2

VN Clase 1= VN1+VN2+VN3+VN4

Figura 43

Obtención de Verdaderos Negativos de la clase 2 de la Matriz de confusión de 3 clases.

Obtención de Verdaderos Negativos Clase 2

Clase 0	VN1	VN2	
Clase 1	VN3	VN4	
Clase 2			VP2
	Clase 0	Clase 1	Clase 2

VN Clase 2= VN1+VN2+VN3+VN4

Obtención de Falsos Positivos

Para el caso de obtención de Falsos Positivos (FP) , el procedimiento que se realiza es, tomar como referencia el VP de la clase requerida, y se suman los valores pertenecientes a la columna en la que se encuentra dicho VP, sin tomar en cuenta para la suma el valor de VP. En la Figura 44 se observa de mejor manera dicha explicación.

Figura 44

Obtención de Falsos Positivos de la Matriz de confusión de 3 clases.

Obtención de Falsos Positivos

Clase 0	VP0	FP3	FP5
Clase 1	FP1	VP1	FP6
Clase 2	FP2	FP4	VP2
	Clase 0	Clase 1	Clase 2

FP Clase 0= FP1+FP2
FP Clase 1= FP3+FP4
FP Clase 0= FP5+FP6

Obtención de Falsos Negativos

Para conseguir los denominados Falsos Negativos (FN), se toma un procedimiento parecido al de los FP, a diferencia de que, en este caso, se toma como referencia el VP de la clase requerida, y se suman los valores pertenecientes a la fila en la que se encuentra dicho VP, sin tomar para la suma el valor de VP. La Figura 45 muestra de mejor manera este cálculo.

Figura 45

Obtención de Falsos Negativos de la Matriz de confusión de 3 clases.

Obtención de Falsos Negativos

Clase 0	**VP0**	**FN1**	**FN2**
Clase 1	**FN3**	**VP1**	**FN4**
Clase 2	**FN5**	**FN6**	**VP2**
	Clase 0	**Clase 1**	**Clase 2**

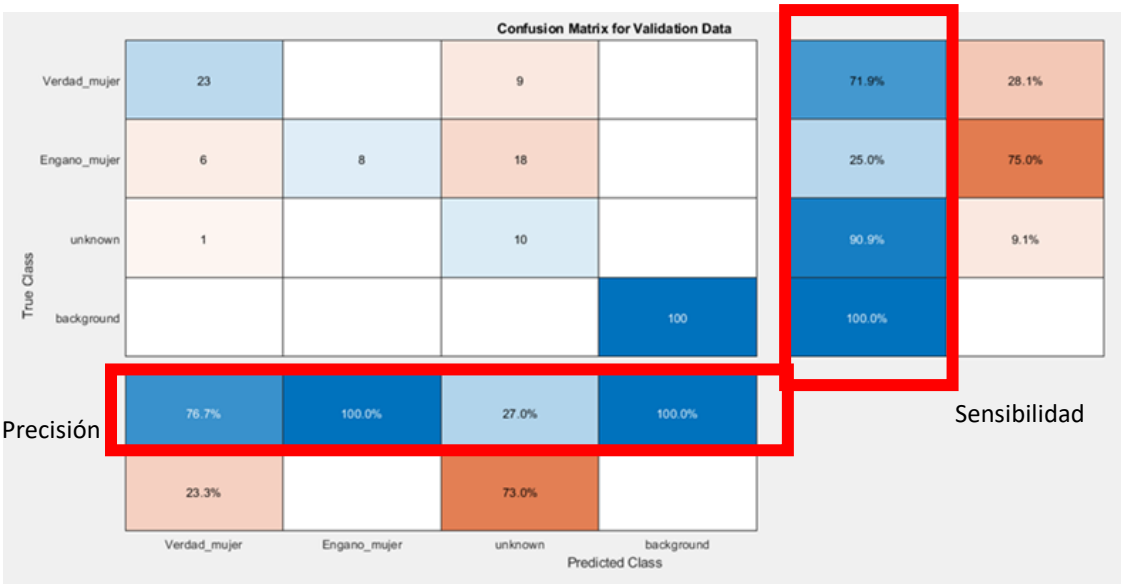
FN Clase 0=	FN1+FN2
FN Clase 1=	FN3+FN4
FN Clase 0=	FN5+FN6

Bajo esta explicación, en el ejemplo de matriz de confusión de la Figura 39, se denota que, nuestros Verdaderos Positivos (VP) son aquellos resultados obtenidos en color celeste, en la diagonal de la matriz, de cada clase, es decir, si queremos los Verdaderos Positivos de la clase Verdad_mujer, claramente son 23 del total de 32 señales, mientras que los VP de la clase Engano_mujer son 8 del total de 32.

Adicionalmente, se tiene también, valores que se muestran en porcentajes, estos corresponden a la sensibilidad (columna) y precisión (fila), como se evidencia en la Figura 46.

Figura 46

Ejemplo de resultado de Matriz de Confusión (Mujeres) - Sensibilidad y Precisión



Estos valores corresponden al cálculo de los parámetros “precisión” y “sensibilidad” que se obtienen a partir de las ecuaciones 8 y 9.

Precisión (P)

$$P(\%) = \frac{VP}{VP + FP} \times 100 \qquad (8)$$

Sensibilidad (R)

$$R(\%) = \frac{VP}{VP + FN} \times 100 \qquad (9)$$

Se aplica el procedimiento de obtención de los VP, FP, y FN, necesarios para reemplazar dichos valores en la ecuación, por cada clase deseada, teniendo como resultado para la clase Verdad_mujer los siguientes reemplazos:

$$\boldsymbol{P(\%)_{Verdadmujer}} = \frac{\mathbf{23}}{\mathbf{23 + (6 + 1)}} \times \mathbf{100} = \frac{\mathbf{23}}{\mathbf{30}} \mathbf{\ x\ 100} = \mathbf{76,7\%}$$

$$\mathbf{R(\%)_{Verdadmujer}} = \frac{\mathbf{23}}{\mathbf{23 + (9 + 0)}} \times \mathbf{100} = \frac{\mathbf{23}}{\mathbf{32}}\ \mathbf{x\ 100} = \mathbf{71,9\%}$$

Así se observa en la Figura 47, dichos resultados:

Figura 47

Resultados de Sensibilidad y Precisión de la Clase Verdad_mujer en la matriz de confusión.

De la misma manera, se aplica a la clase Engano_mujer, obteniendo:

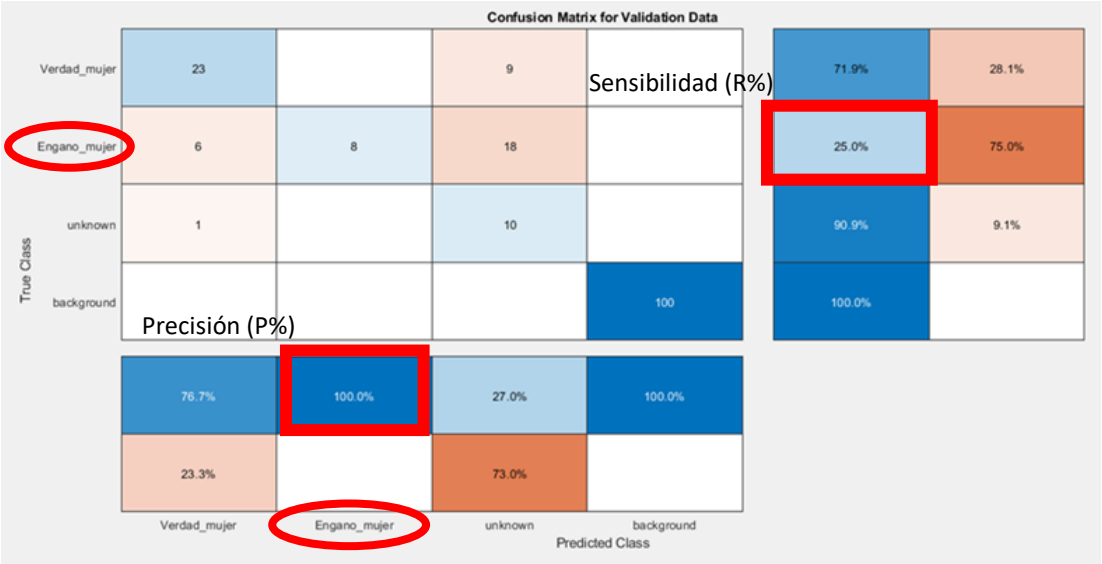
$$\boldsymbol{P(\%)_{Engañomujer}} = \frac{\mathbf{8}}{\mathbf{8 + (0 + 0)}} \times \mathbf{100} = \frac{\mathbf{8}}{\mathbf{8}}\ \mathbf{x\ 100} = \mathbf{100\%}$$

$$\mathbf{R(\%)_{Engañomujer}} = \frac{\mathbf{8}}{\mathbf{8 + (18 + 6)}} \times \mathbf{100} = \frac{\mathbf{8}}{\mathbf{32}}\ \mathbf{x\ 100} = \mathbf{25\%}$$

Los resultados se observan en la Figura 48.

Figura 48

Resultados de Sensibilidad y Precisión de la Clase Engano_mujer en la matriz de confusión.



Este mismo procedimiento se aplica a hombres, con cada modelo mencionado, denotando así la forma en cómo se generó el entrenamiento de las CNN para poder obtener los resultados que se evidencian y detallan en el capítulo 4.

Con la explicación antes detallada, se enlista una muestra general de todos los resultados obtenidos del entrenamiento como se observa en las Figuras 49 a la 56 para hombres y de las Figuras 57 a la 64 para mujeres, de los cuales, su respectivo análisis y cálculo se evidencian en el capítulo 4.

Muestra General de los Experimentos del Entrenamiento de la Red Neuronal (Hombres).

Figura 49

Experimento 1-Entrenamiento Red Neuronal (Hombres)

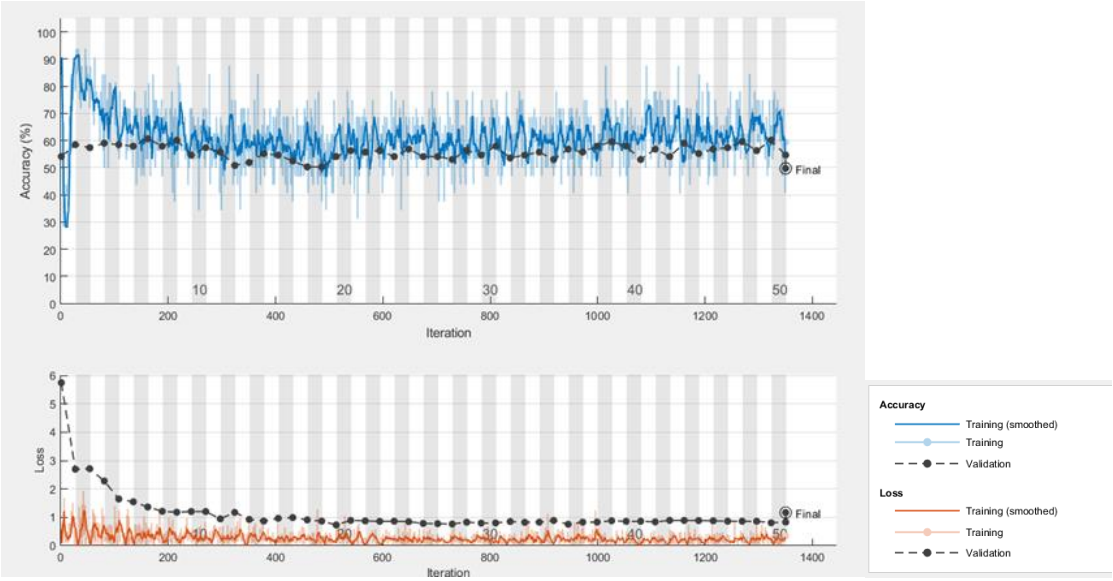


Figura 50

Experimento 1- Matriz de Confusión (Hombres)

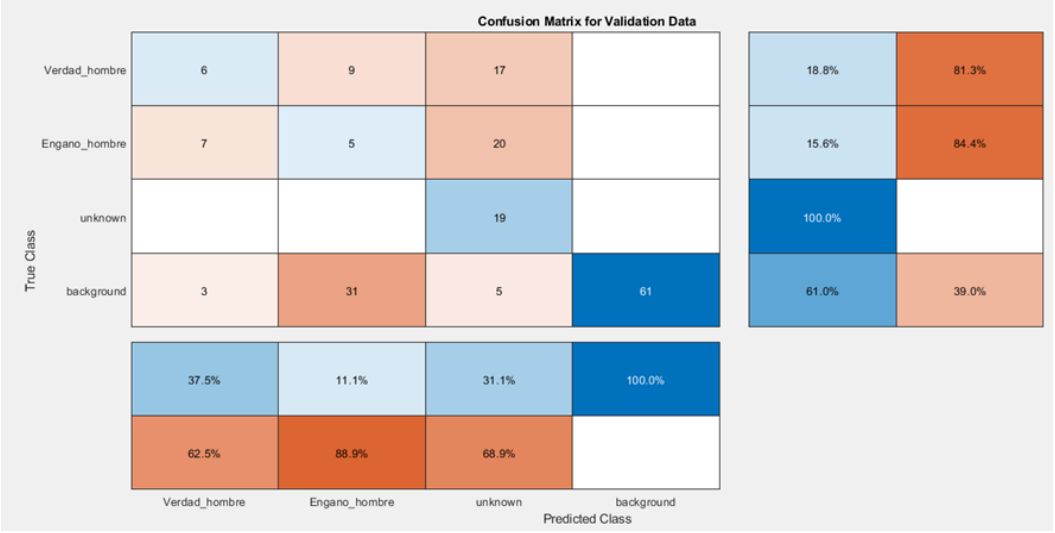


Figura 51

Experimento 2-Entrenamiento Red Neuronal (Hombres)

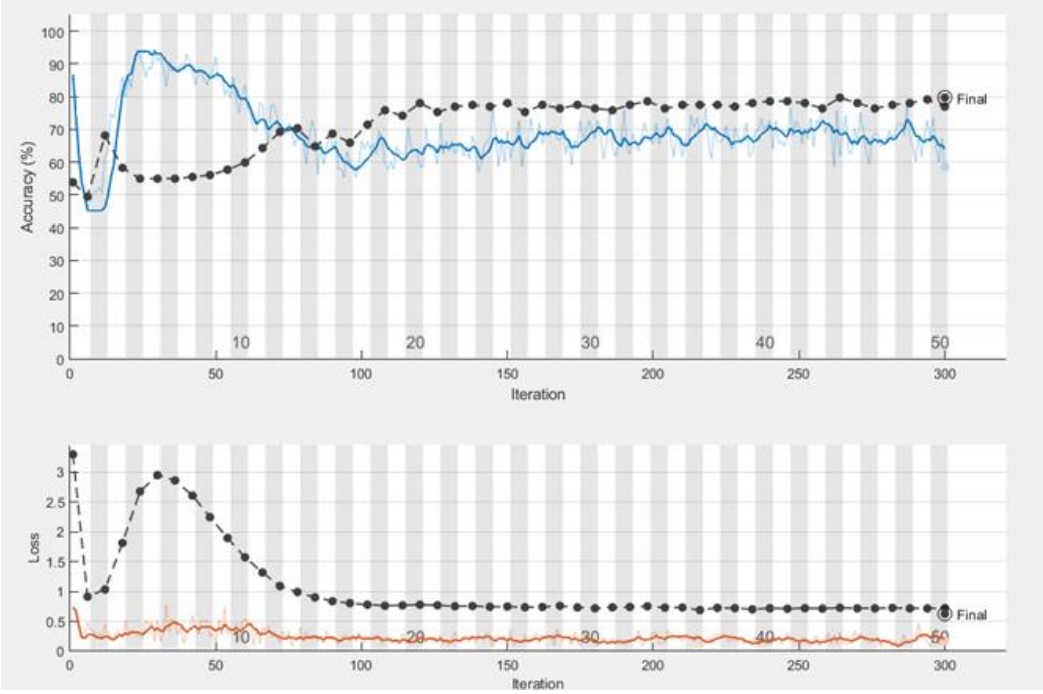


Figura 52

Experimento 2- Matriz de Confusión (Hombres)

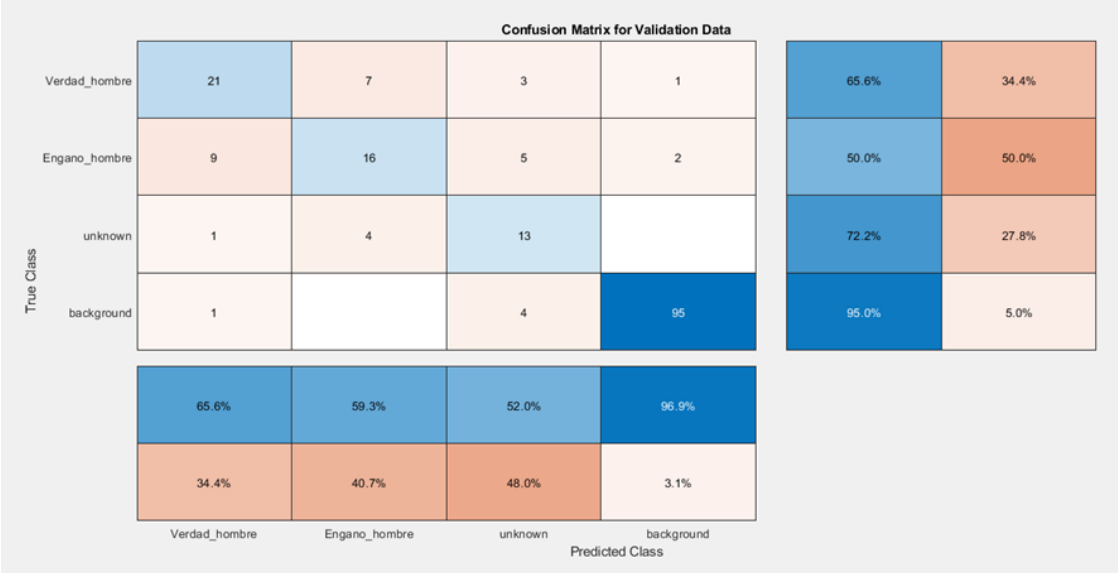


Figura 53

Experimento 3-Entrenamiento Red Neuronal (Hombres)

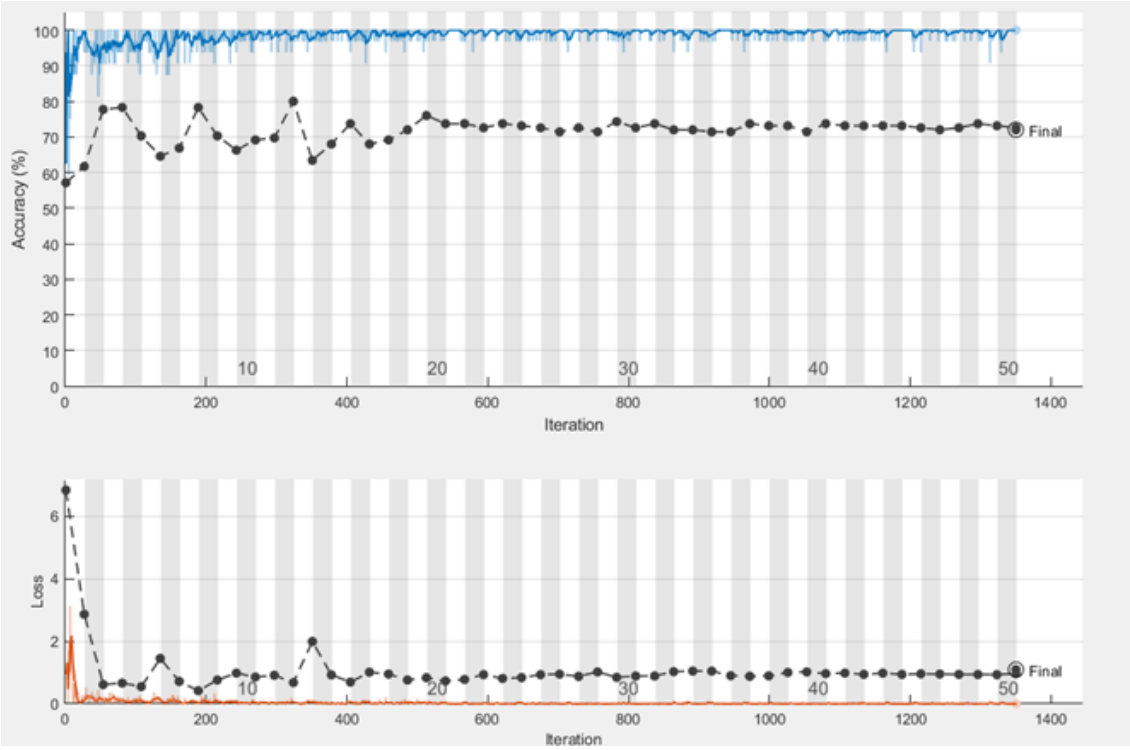


Figura 54

Experimento 3- Matriz de Confusión (Hombres)

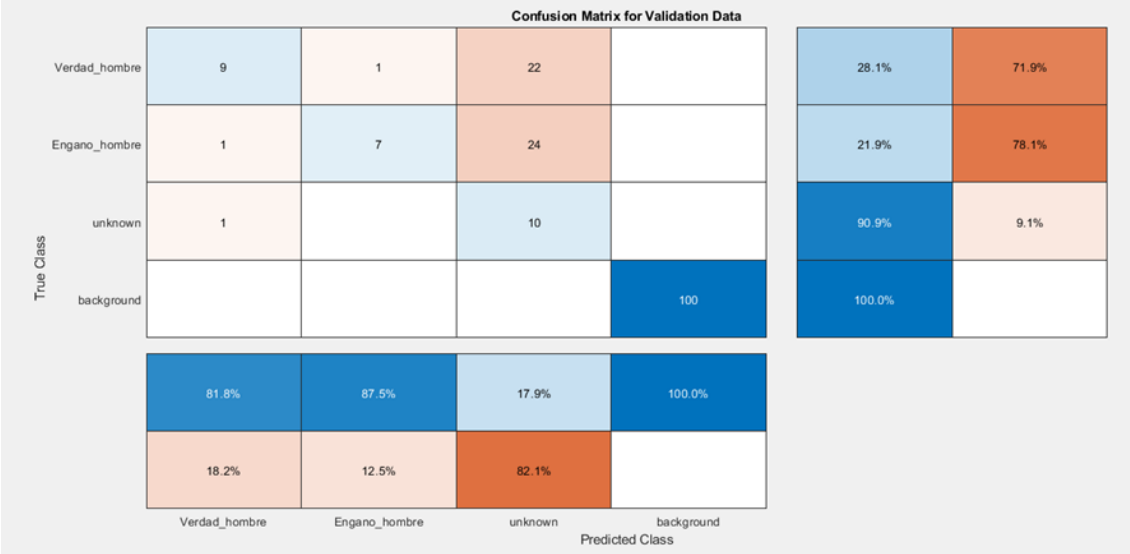


Figura 55

Experimento 4-Entrenamiento Red Neuronal (Hombres)

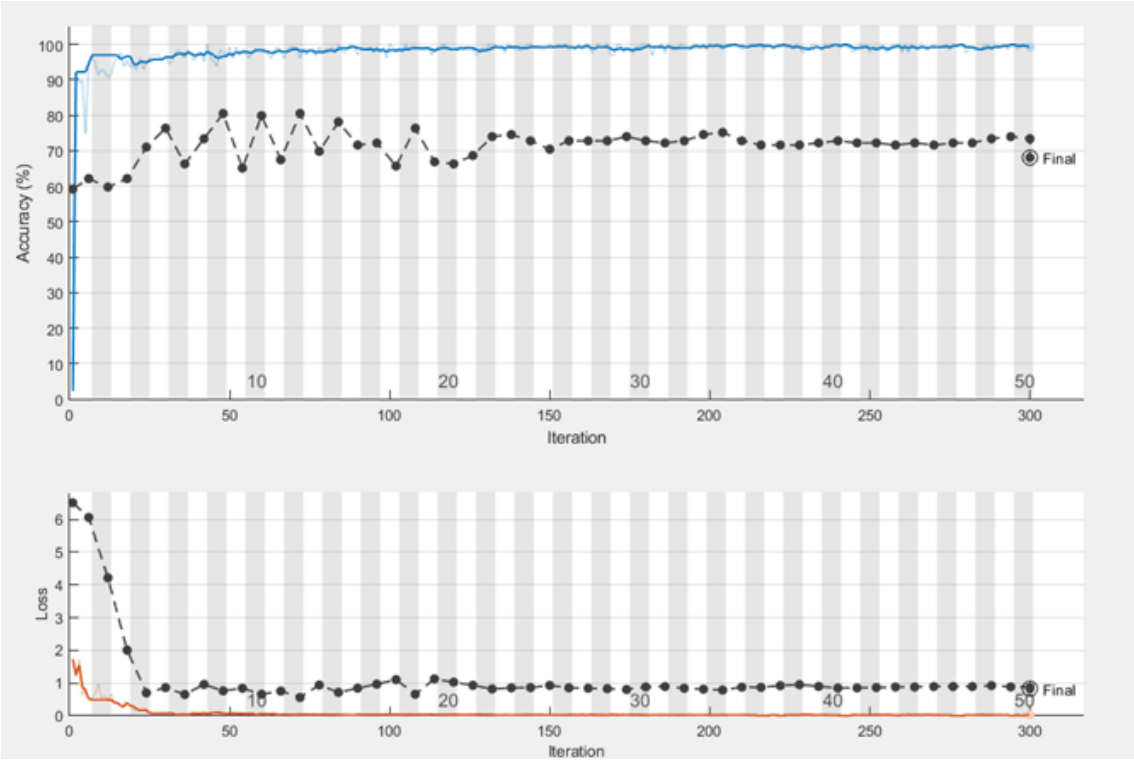
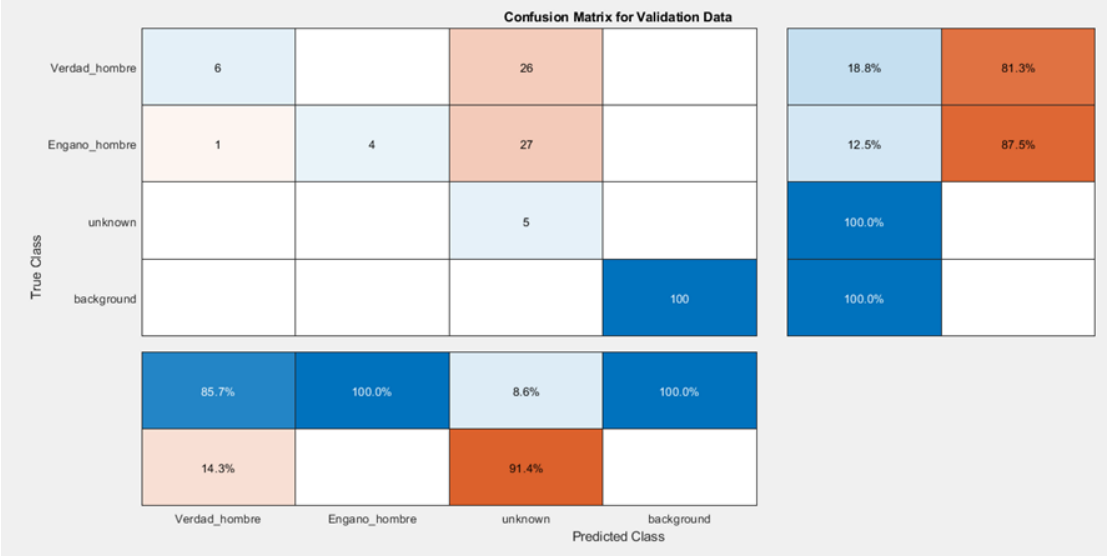


Figura 56

Experimento 4- Matriz de Confusión (Verdad)



Muestra General de los Experimentos del Entrenamiento de la Red Neuronal (Mujeres).

Figura 57

Experimento 1-Entrenamiento Red Neuronal (Mujeres)

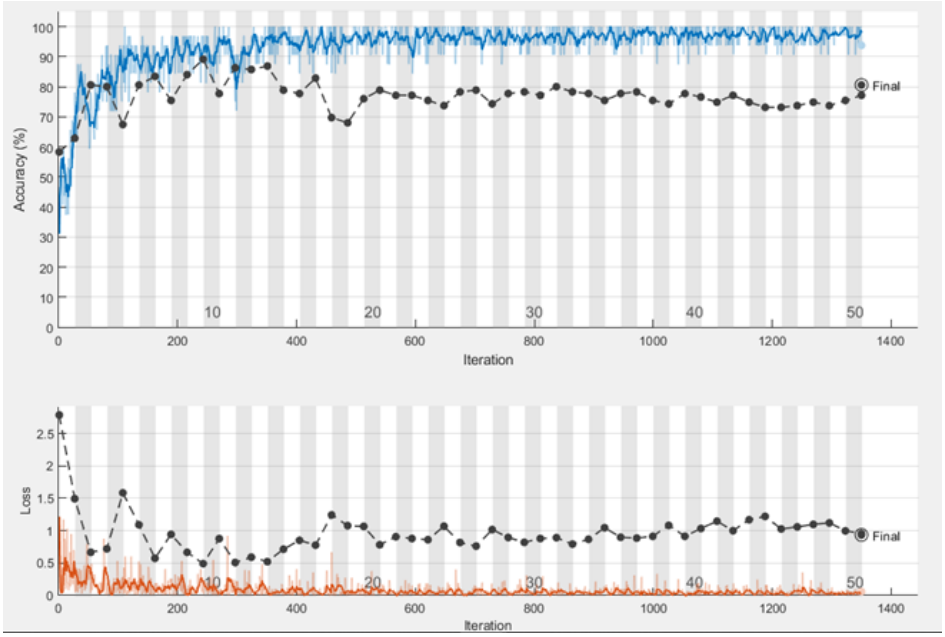


Figura 58

Experimento 1- Matriz de Confusión (Mujeres)

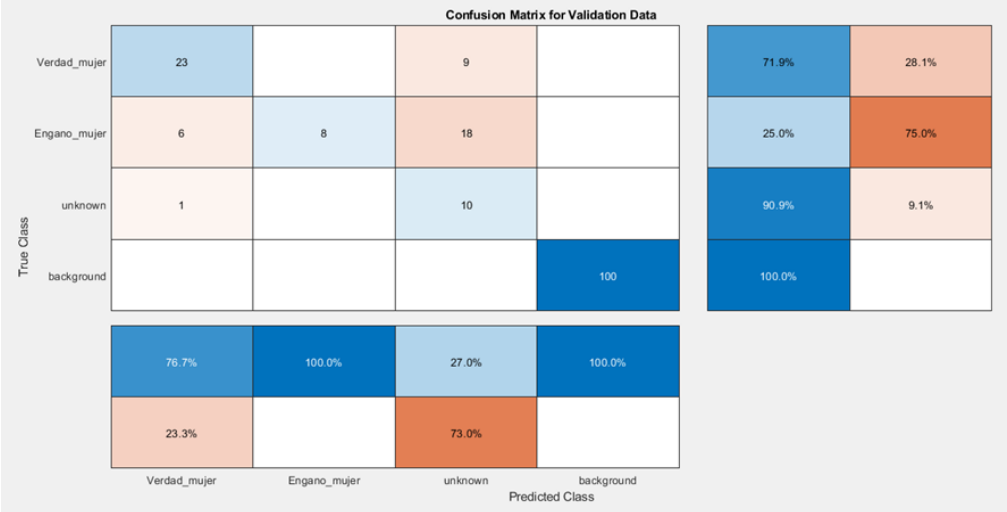


Figura 59

Experimento 2-Entrenamiento Red Neuronal (Mujeres)

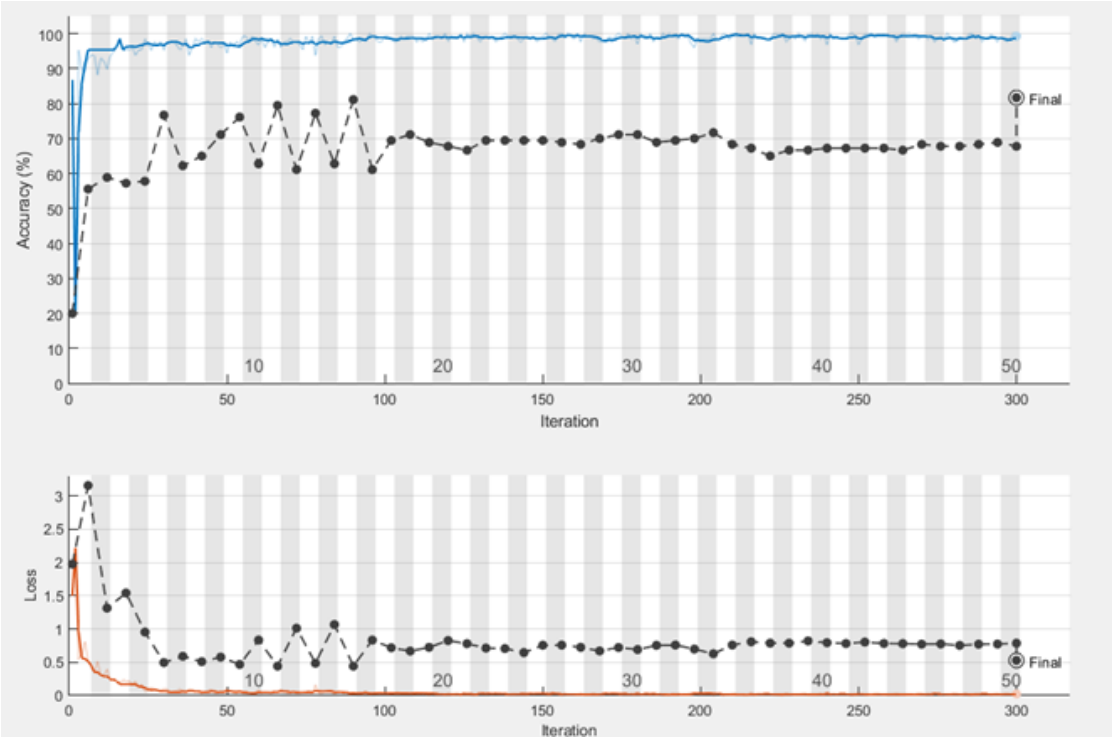


Figura 60

Experimento 2- Matriz de Confusión (Mujeres)

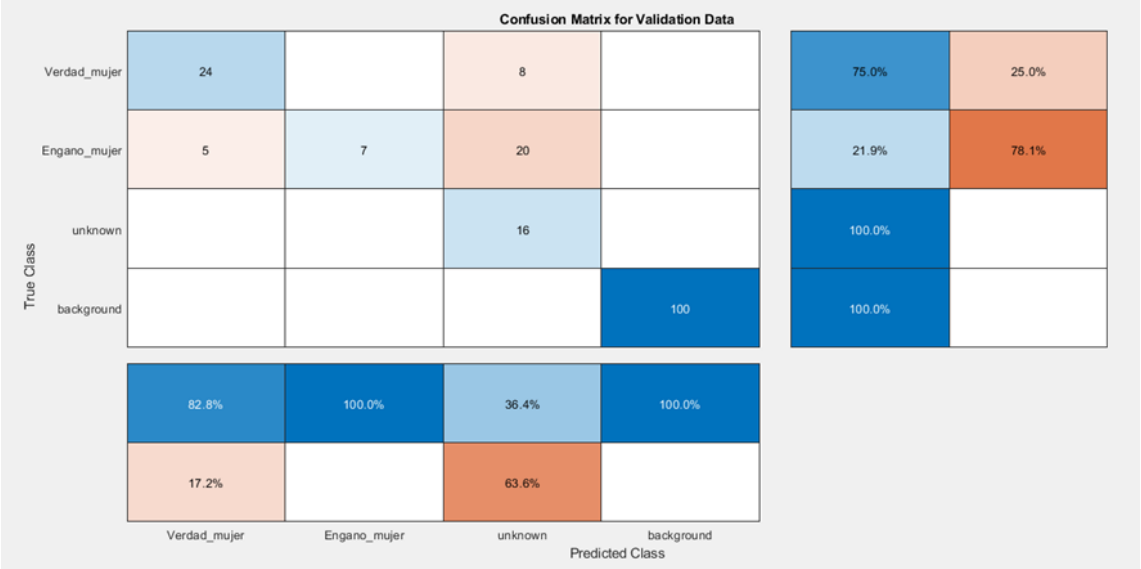


Figura 61

Experimento 3-Entrenamiento Red Neuronal (Mujeres)

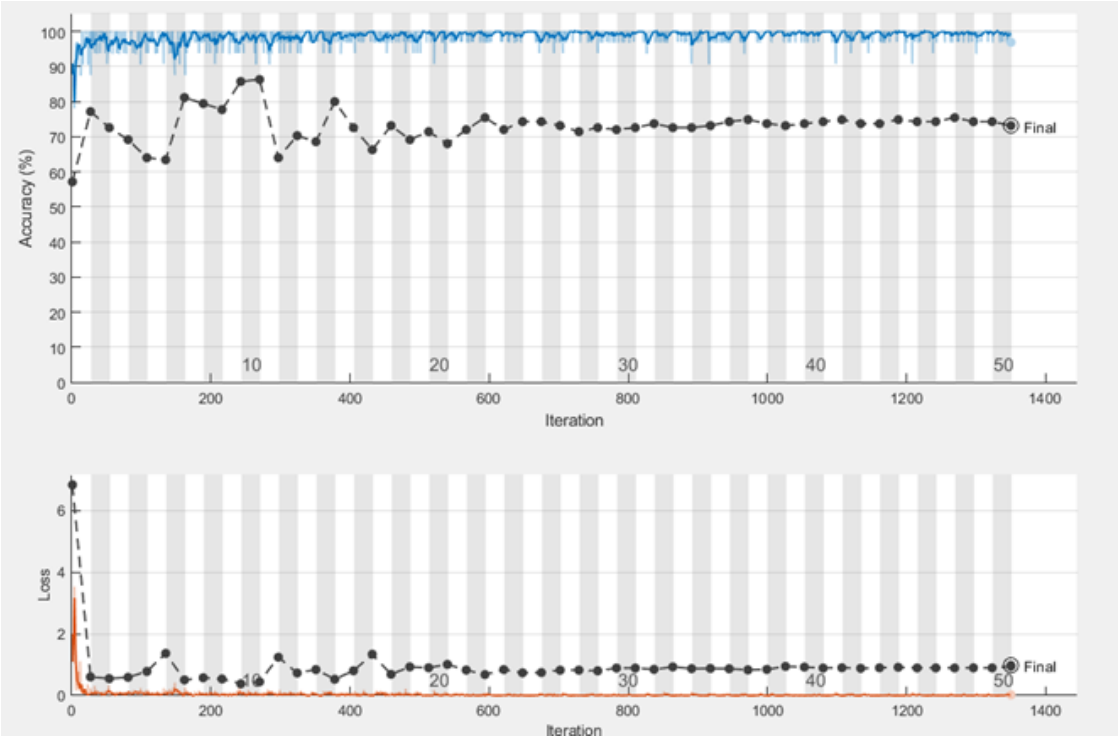


Figura 62

Experimento 3- Matriz de Confusión (Mujeres)

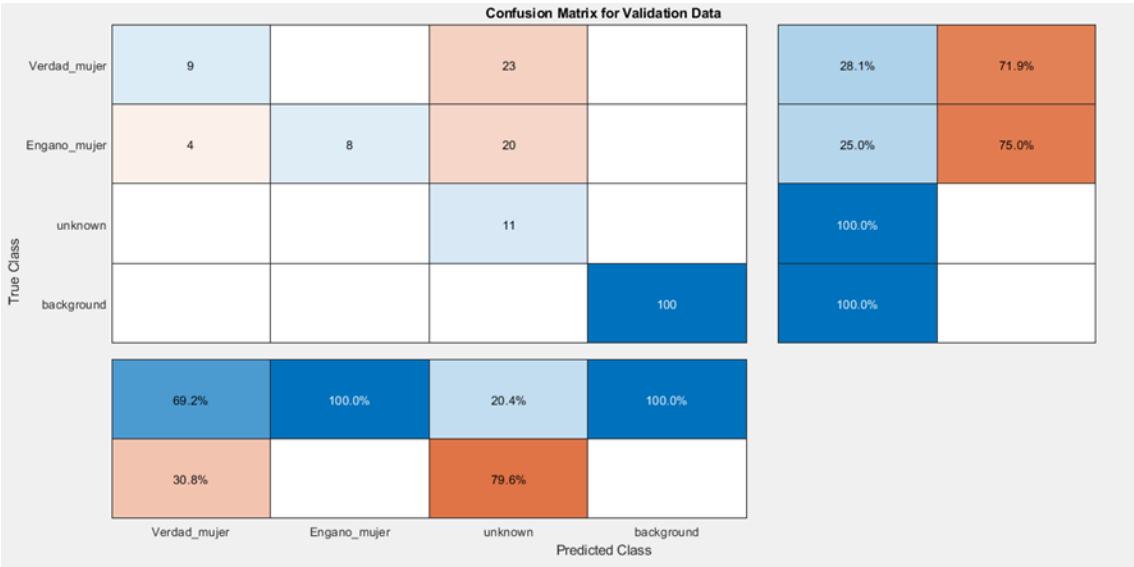


Figura 63

Experimento 4-Entrenamiento Red Neuronal (Mujeres)

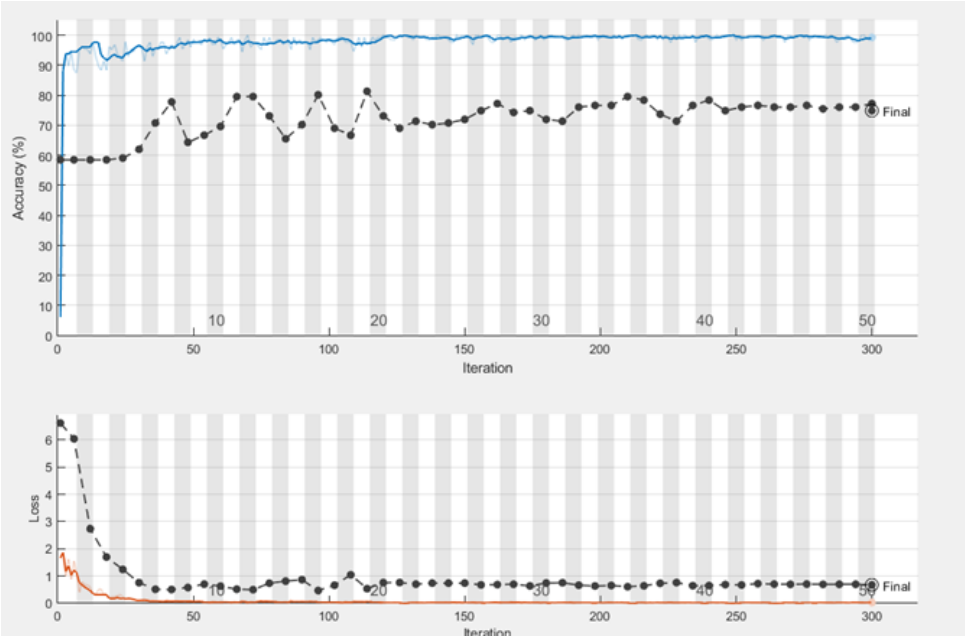
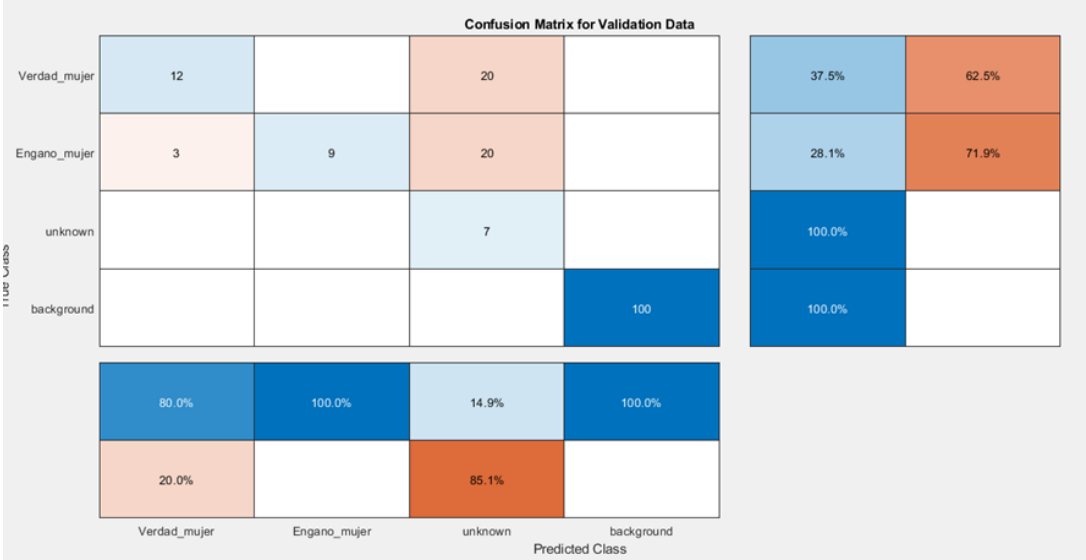


Figura 64

Experimento 4- Matriz de Confusión (Mujeres)



Capitulo IV

Resultados de Investigación

Análisis de Resultados

En el presente capítulo se muestra el rendimiento del sistema al ejecutarse sobre la base de datos mencionada anteriormente en el capítulo número 3. Se evalúa por géneros masculino y femenino por separado para evitar conflictos en el análisis debido a que, en muchas ocasiones, las señales de engaño en un hombre se encontraban parecidas o con valores similares en amplitud al de las verdades de una mujer. Los resultados a evaluar se miden en base a los siguientes parámetros:

- Verdaderos positivos (VP)
- Falsos positivos (FP)
- Verdaderos negativos (VN)
- Falsos negativos (FN)

Bajo estos parámetros se analiza la exactitud, precisión, sensibilidad y especificidad del sistema, para lo cual, se utilizan las ecuaciones 10, 11, 12 y 13, respectivamente.

Exactitud (A)

$$A(\%) = \frac{VP + VN}{VP + VN + FP + FN} \times 100 \quad (10)$$

Precisión (P)

$$P(\%) = \frac{VP}{VP + FP} \times 100 \quad (11)$$

Sensibilidad (R)

$$R(\%) = \frac{VP}{VP + FN} \times 100 \quad (12)$$

Especificidad (S)

$$S(\%) = \frac{VN}{VN + FP} \times 100 \quad (13)$$

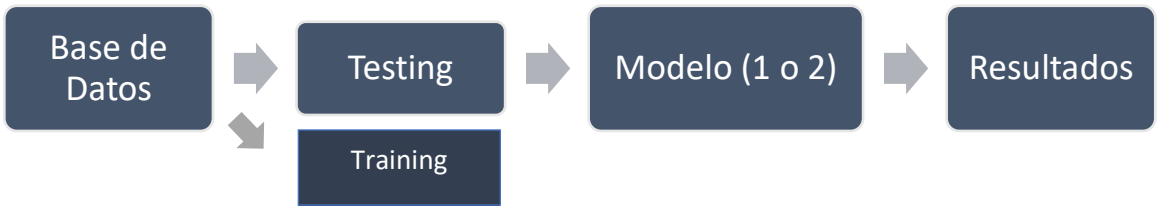
Estas métricas permiten analizar el desempeño del algoritmo de detección, las cuales se definen dentro del presente proyecto de investigación como:

- Exactitud (A%): Es la cantidad de predicciones positivas que fueron clasificadas correctamente, del total de casos examinados.
- Precisión (P%): En forma práctica es el porcentaje total de señales detectados correctamente.
- Sensibilidad (R%): Es la probabilidad de clasificar correctamente a las señales de verdad, como verdad.
- Especificidad (S%): Es la probabilidad de clasificar correctamente a una señal de engaño, como engaño.

La Figura 65, muestra un esquema general a seguirse para los experimentos.

Figura 65

Esquema General del Proceso de Pruebas



Nota. Elaboración propia.

Análisis del Desempeño del Sistema de Detección de Verdades y Engaños (Hombres)

Evaluación del Experimento 1

El experimento 1 realiza la medición del desempeño del sistema con las señales de entrada de la base de datos hombre, una utilización de 10 capas convolucionales y un optimizador de Adam de tamaño 32. Los resultados que otorga la red neuronal muestran los siguientes valores:

- Error de entrenamiento: 3.7457 %
- Error de validación: 5.0273 %

El error de entrenamiento nos ayuda a identificar si el modelo utilizado es el adecuado, si este porcentaje es alto, significa que el modelo no puede aprender la relación entre los datos de entrada y los resultados, es decir, el modelo tiene problemas para aprender.

El error de validación nos permite identificar si el modelo utilizado ha aprendido el patrón que hay entre los datos de entrada y los resultados, si es alto, aprende de "memoria", si es bajo, aprendió el patrón.

Tabla 2

Desempeño del Sistema de Detección de Verdades y Engaños (Hombres) - Experimento 1

	Exactitud A(%)	Precisión P(%)	Sensibilidad R(%)	Especificidad S(%)
Verdad Hombre	60.24	46.15	18.80	86.27
Engaño Hombre	56.63	35.71	15.63	82.35

Con los valores de la Tabla 2 se muestran los siguientes resultados obtenidos: exactitud de 58.44 %, precisión 40.93 %, sensibilidad 17.22 % y especificidad de 84.31 %.

Evaluación del Experimento 2

El experimento 2, es una prueba muy parecida al experimento 1, realizando la medición del desempeño del sistema con la utilización de 10 capas convolucionales, pero con la diferencia del empleo de un optimizador de Adam de tamaño 128. Los resultados que otorga la red neuronal muestran los siguientes valores:

- Error de entrenamiento: 1.45 %
- Error de validación: 2.0329 %

Tabla 3

Desempeño del Sistema de Detección de Verdades y Engaños (Hombres) - Experimento 2

	Exactitud A(%)	Precisión P(%)	Sensibilidad R(%)	Especificidad S(%)
Verdad Hombre	74.39	67.74	65.63	80.00
Engaño Hombre	67.07	59.26	50.00	78.00

Con los valores de la Tabla 3 se muestran los siguientes resultados obtenidos: exactitud de 70.73 %, precisión 63.5 %, sensibilidad 57.82 % y especificidad de 79.00 %.

Evaluación del Experimento 3

El experimento 3, realiza la medición del desempeño del sistema, la diferencia respecto a los experimentos 1 y 2, es que utiliza 5 capas convolucionales y un optimizador de Adam de tamaño 32. Los resultados que otorga la red neuronal muestran los siguientes valores:

- Error de entrenamiento: 1.75714 %
- Error de validación: 2.8 %

Tabla 4

Desempeño del Sistema de Detección de Verdades y Engaños (Hombres) – Experimento 3

	Exactitud A(%)	Precisión P(%)	Sensibilidad R(%)	Especificidad S(%)
Verdad Hombre	66.67	81.82	28.13	95.35
Engaño Hombre	65.33	87.50	21.88	97.67

Con los valores de la Tabla 4 se muestran los siguientes resultados obtenidos: exactitud de 66.00 %, precisión 84.66 %, sensibilidad 25.01 % y especificidad de 96.51 %.

Evaluación del Experimento 4

El experimento 4 realiza la medición del desempeño del sistema de manera parecida al experimento 3 con la utilización de 5 capas convolucionales, no obstante, el presente experimento utiliza un optimizador de Adam de tamaño 128. Los resultados que otorga la red neuronal muestran los siguientes valores:

- Error de entrenamiento: 1.56754 %
- Error de validación: 3.1952 %

Tabla 5

Desempeño del Sistema de Detección de Verdades y Engaños (Hombres) - Experimento 4

	Exactitud A(%)	Precisión P(%)	Sensibilidad R(%)	Especificidad S(%)
Verdad Hombre	60.87	85.71	18.80	97.30
Engaño Hombre	59.42	100.00	12.50	100.00

Con los valores de la Tabla 5 se muestran los siguientes resultados obtenidos: exactitud de 60.15 %, precisión 92.86 %, sensibilidad 15.65 % y especificidad de 98.65 %.

Análisis del desempeño del sistema de detección de verdades y engaños (Mujeres)

Evaluación del Experimento 1

El experimento 1 realiza la medición del desempeño del sistema con la utilización de 10 capas convolucionales y un optimizador de Adam de tamaño 32. Los resultados que otorga la red neuronal muestran los siguientes valores:

- Error de entrenamiento: 1.4623 %
- Error de validación: 1.9428 %

Tabla 6

Desempeño del Sistema de Detección de Engaños y Verdades (Mujeres) - Experimento 1

	Exactitud A(%)	Precisión P(%)	Sensibilidad R(%)	Especificidad S(%)
Verdad Mujer	78.67	76.67	71.90	83.72
Engaño Mujer	68.00	100.00	25.00	100.00

Con los valores de la Tabla 6 se muestran los siguientes resultados obtenidos: exactitud de 73.34 %, precisión 88.34 %, sensibilidad 48.45 % y especificidad de 91.86 %.

Evaluación del Experimento 2

El experimento 2 realiza la medición del desempeño del sistema con la utilización de 10 capas convolucionales y un optimizador de Adam de tamaño 128. Los resultados que otorga la red neuronal muestran los siguientes valores:

- Error de entrenamiento: 0.9142 %
- Error de validación: 1.8333 %

Tabla 7

Desempeño del Sistema de Detección de Engaños y Verdades (Mujeres) - Experimento 2

	Exactitud A(%)	Precisión P(%)	Sensibilidad R(%)	Especificidad S(%)
Verdad Mujer	83.75	82.76	75.00	89.58
Engaño Mujer	68.75	100.00	21.88	100.00

Con los valores de la Tabla 7 se muestran los siguientes resultados obtenidos: exactitud de 76.25 %, precisión 91.38 %, sensibilidad 48.44 % y especificidad de 94.79 %.

Evaluación del Experimento 3

El experimento 3 realiza la medición del desempeño del sistema con la utilización de 5 capas convolucionales y un optimizador de Adam de tamaño 32. Los resultados que otorga la red neuronal muestran los siguientes valores:

- Error de entrenamiento: 0.68571 %
- Error de validación: 2.6857 %

Tabla 8

Desempeño del Sistema de Detección de Engaños y Verdades (Mujeres) - Experimento 3

	Exactitud A(%)	Precisión P(%)	Sensibilidad R(%)	Especificidad S(%)
Verdad Mujer	64.00	69.23	28.13	90.70
Engaño Mujer	68.00	100.00	25.00	100.00

Con los valores de la Tabla 8 se muestran los siguientes resultados obtenidos: exactitud de 66.00 %, precisión 84.62 %, sensibilidad 26.57 % y especificidad de 95.35 %.

Evaluación del Experimento 4

El experimento 4 realiza la medición del desempeño del sistema con la utilización de 5 capas convolucionales y un optimizador de Adam de tamaño 128. Los resultados que otorga la red neuronal muestran los siguientes valores:

- Error de entrenamiento: 0.6841 %
- Error de validación: 2.5146 %

Tabla 9

Desempeño del Sistema de Detección de Engaños y Verdades (Mujeres) - Experimento 4

	Exactitud A(%)	Precisión P(%)	Sensibilidad R(%)	Especificidad S(%)
Verdad Mujer	67.61	80.00	37.50	92.31
Engaño Mujer	67.61	100.00	28.13	100.00

Con los valores de la Tabla 9 se muestran los siguientes resultados obtenidos: exactitud de 67.61 %, precisión 90.00 %, sensibilidad 32.82 % y especificidad de 96.16 %.

Análisis Total de los Experimentos

El resumen de los resultados obtenidos por la predicción en cada uno de los experimentos se muestra en la Tabla 10, Tabla 11, Figura 66 y Figura 67 respectivamente.

Tabla 10

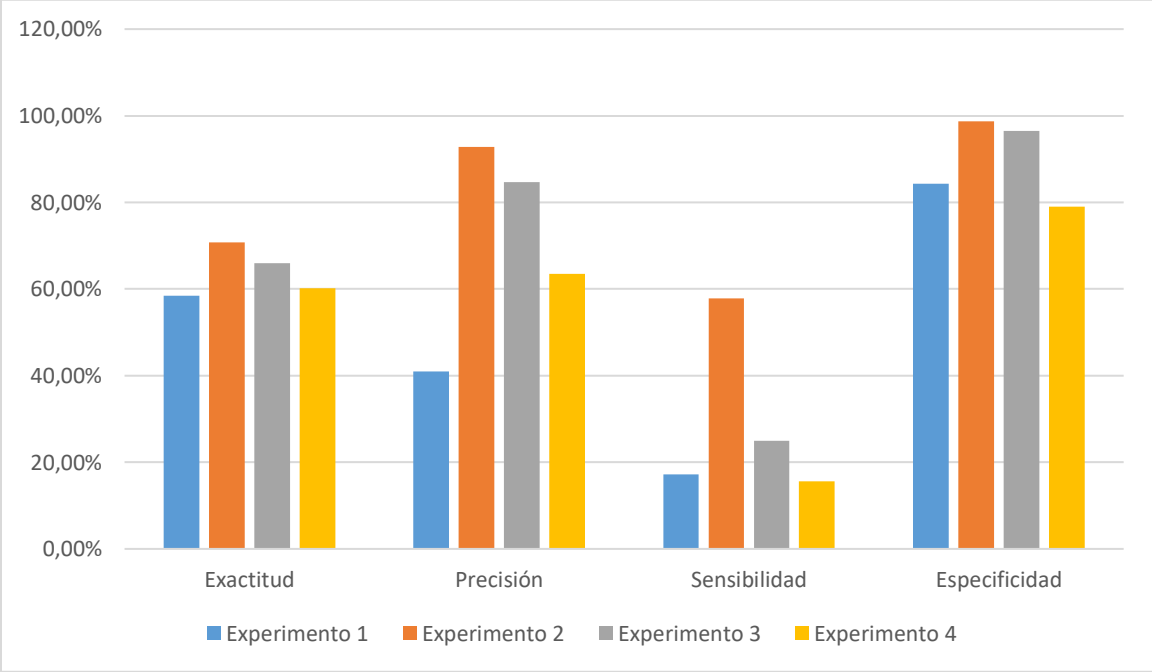
Desempeño del Sistema de Detección de Verdades y Engaños (Hombres)

	Exactitud A(%)	Precisión P(%)	Sensibilidad R(%)	Especificidad S(%)	Error de Entrenamiento	Error de Validación
Experimento 1	58.44 %	40.93 %	17.22 %	84.31 %	3.7457%	5.0273%
Experimento 2	**70.73 %**	**92.86 %**	**57.82 %**	**98.65 %**	**1.45 %**	**2.0329%**
Experimento 3	66.00 %	84.66 %	25.01 %	96.51 %	1.75714%	2.8%
Experimento 4	60.15 %	63.50 %	15.65 %	79.00 %	1.56754%	3.1952%

De la Tabla 10 se denota, que el experimento 2 (10 capas, Adam 128), obtuvo los mejores resultados en porcentajes de Exactitud, Precisión, Sensibilidad y Especificidad respecto al resto de experimentos que utilizan menor cantidad de capas y un optimizador más bajo; de manera análoga, se observa que el error de entrenamiento es el más bajo en el experimento 2, al igual que el error de validación, si nos fijamos con respecto al experimento 1 (en el cual el número de capas es menor, y el optimizador de Adam es de 32), dichos errores disminuyeron.

Figura 66

Desempeño del Sistema de Detección de Verdades y Engaños (Hombres)



De acuerdo a los resultados mostrados por la Figura 66, se denota mejores resultados del experimento 2 del sistema de detección de verdades y engaños (Hombres) en el análisis de los parámetros precisión, sensibilidad, y cercano a los demás experimentos respecto a la exactitud y la especificidad. Experimento en el que se realizó al aumentar el tamaño del optimizador y el número de capas convolucionales. Dentro del análisis de errores se obtuvo de igual manera el valor más bajo en el conjunto de error de entrenamiento y validación, lo cual

nos indica que el modelo fue el adecuado y aprendió de manera correcta. Al aumentar el tamaño de capas convolucionales, ayuda a que se extraigan características más representativas, por lo que, a mayor cantidad de capas, mejor resultado se obtiene de nuestra CNN.

Tabla 11

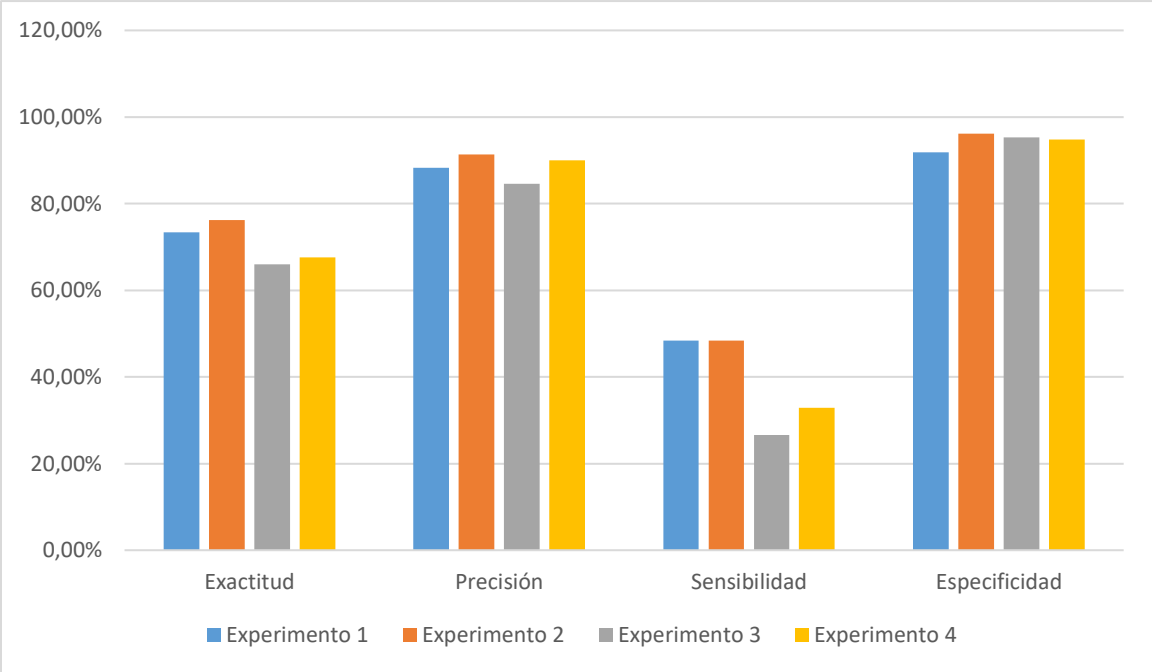
Desempeño del Sistema de Detección de Verdades y Engaños (Mujeres)

	Exactitud A(%)	Precisión P(%)	Sensibilidad R(%)	Especificidad S(%)	Error de Entrenamiento	Error de Validación
Experimento 1	73.34 %	88.34 %	48.44 %	91.86 %	1.4623%	1.9428%
Experimento 2	**76.25 %**	**91.38 %**	**48.45 %**	**96.16%**	**0.9142%**	**1.8333%**
Experimento 3	66.00 %	84.62 %	26.57 %	95.35 %	0.68571%	2.6857%
Experimento 4	67.61 %	90.00 %	32.82 %	94.79 %	0.6841%	2.5146%

De la Tabla 11 se denota, que el experimento 2 (10 capas, Adam 128), obtuvo los mejores resultados en porcentajes de Exactitud, Precisión, Sensibilidad y Especificidad; de manera análoga, se observa que el error de entrenamiento no es el más bajo en el experimento 2, sin embargo, observando el error de validación tendríamos el conjunto de ambos errores más bajos; si nos fijamos con respecto al experimento 1 (en el cual el número de capas es menor, y el optimizador de Adam es de 32), dichos errores disminuyeron.

Figura 67

Desempeño del Sistema de Detección de Verdades y Engaños (Mujeres)



De acuerdo a los resultados mostrados por la Figura 67 se observa un mejor desempeño del experimento 2 del sistema de detección de engaños y verdades en Mujeres en el análisis de los parámetros exactitud, precisión y especificidad, pero muy cercano al experimento 1 respecto a la sensibilidad. De igual manera, dicho experimento expone mejores resultados frente al resto, de acuerdo a que se realizó el aumento del tamaño de la ventana y el número de capas convolucionales. Los errores de entrenamiento y validación de igual manera son bajos, lo cual nos asegura que el modelo utilizado fue el más adecuado y aprendió de manera correcta.

Al aumentar el tamaño de capas convolucionales, ayuda a que se extraigan características más representativas, por lo que, a mayor cantidad de capas, mejor resultado se obtiene de nuestra CNN.

Después de las pruebas realizadas se denota que es posible detectar engaños y verdades a partir de señales de voz de hombres y mujeres con el método y modelos seleccionados en el presente trabajo de investigación, y los resultados mostrados demuestran la validez necesaria para confiar en que lo aquí expuesto servirá como base para obtención de mejoras en investigaciones futuras.

Capítulo V

Conclusiones y Recomendaciones

Conclusiones

La utilización de Redes Neuronales Convolucionales para la predicción de verdades y engaños a partir de señales de voz en hombres y mujeres, es un método válido y confiable de aprendizaje en Deep Learning.

Se evitó realizar el entrenamiento uniendo señales de voz de hombres con las de mujeres, debido a que la CNN puede entrenarse, y aprender de manera errónea confundiendo la predicción de hombres con la de mujeres.

El mejor método para la detección de engaños y verdades en hombres y en mujeres, fue el de utilización de una CNN con optimizador de Adam de 128 y 10 capas convolucionales en la red.

El experimento 2, en ambos casos, presenta mejores resultados con relación al resto de experimentos tanto de hombres como de mujeres gracias a la variación de los valores del optimizador de Adam entre 32 y 128, concluyendo que el que entregó los mejores resultados tanto en hombres y mujeres es 128, ayudando a ver reflejado dichos resultados en los valores de exactitud, precisión, sensibilidad y especificidad llegando muy cercanamente a un 100%.

Las variaciones de capas convolucionales y el optimizador de Adam ayudaron a obtener valores muy bajos de error de entrenamiento como de validación concluyendo que el modelo utilizado aprendió de manera correcta gracias al aumento de capas convolucionales.

Recomendaciones

Antes de iniciar con los entrenamientos es recomendable revisar las características de cada archivo de audio, tales como, duración en segundos, tamaño del archivo, formato y frecuencia.

Es recomendable aumentar la cantidad de audios dentro de la base de datos, a fin de tener mayor número de señales a ser reconocidas para obtener CNN mejor entrenadas.

Trabajos Futuros

Se propone probar entrenamientos de CNN con mayor cantidad de capas convolucionales para observar mejoras en la predicción de las verdades y engaños.

Teniendo en cuenta que Matlab es un software matemático con un lenguaje de programación propio, para este mismo estudio se puede utilizar el lenguaje de programación Python, que resulta actualmente más versátil dado a su facilidad de programación general con un compendio de bibliotecas de cálculo científico similares a Matlab, y más actualizados.

Realizar un algoritmo de detección de verdades y engaños mediante la utilización de Recurrent Neural Networks (RNN) o Deep Neural Networks (DNN) y comparar cuál de las diferentes redes otorga mejores resultados en comparación con CNN.

Probar la base de datos con la utilización del framework Tensorflow para Deep Learning, y obtener resultados a fin de comparar el más eficiente.

Generar un aplicativo móvil, a fin de que se pueda utilizar el método de detección en tiempo real de la voz ingresada a través del micrófono de un teléfono móvil y una interfaz gráfica amigable con el usuario.

Referencias Bibliográficas

Acevedo, J. (Noviembre de 2010). *Universidad de Alcalá.* Obtenido de Sistemas Lineales - Análisis de Fourier para Señales : http://agamenon.tsc.uah.es/Asignaturas/ittst/sl/apuntes/Tema4Sesion2.pdf

Ahmadizadeh, M. (2014). An introduction to Short-Time Fourier Transform (STFT). *Sharif University of Technology*.

Araya, M. (2019). Espectrogramas Biocoustica en R.

Arias, S. (2018). Detectar la mentira a través de la voz. *[Universidad de Salamanca.* Trabajo de Pregrado], Salamanca. Obtenido de https://gredos.usal.es/bitstream/handle/10366/138070/TFG_AllAriS_Detectar.pdf?sequence=1&isAllowed=y

Bagnato, J. (2011). Aprende Machine Learning.

Bario, J., García, M. R., Ruiz, I., & Arce, A. (2016). El estrés como respuesta. *Redalyc*, 1(1): 37-48. Obtenido de https://www.redalyc.org/pdf/3498/349832311003.pdf

Beltrán, L. (2003). Simulación de Modelos Ocultos de Markov aplicados al reconocimiento de palabras aisladas, utilizando el programa Matlab. *Escuela de Ingeniería EPN*.

Bravo, S. (2019). IMPLEMENTACIÓN DE UN SISTEMA DE RECONOCIMIENTO AUTOMÁTICO DE ENGAÑOS MEDIANTE EL ANÁLISIS DE LA SEÑAL DE LA VOZ. *Tesis de pregrado].* [Universidad de las Fuerzas Armadas, Quito.

Brownlee , J. (2019). *A Gentle Introduction to Batch Normalization for Deep Neural Networks*. Obtenido de Deep Learning Performance: https://machinelearningmastery.com/batch-normalization-for-training-of-deep-neural-networks/

Campos, C. C., Lopez, E., & Morales, G. (octubre de 2013). *Ciencia uanl.* Recuperado el 2013, de http://cienciauanl.uanl.mx/wp-content/uploads/2014/01/Estudio-cognitivo-de-la-mentira.pdf

Carreño, A. (2017). Detección de sucesos raros con Machine Learning. *[Universidad Politécnica de Madrid.* Trabajo de Posgrado], Madrid. Obtenido de http://oa.upm.es/47931/1/TFM_ANDER_CARRENO_LOPEZ.pdf

Centro Peruano de Audición, Lenguaje y Aprendizaje. (2020). *Unidad de Habla*. Obtenido de CPAL: https://www.cpal.edu.pe/blog/los-efectos-del-estres-y-la-emocion-en-la-voz/

Chi Zaldívar, J. (2012). MEDICIÓN PSICOFISIOLÓGICA: ANALISIS DE ESTRÉS DE VOZ EN CREDIBILIDAD DE TESTIMONIO EN INTERNOS DEL CE.RE.SO.DE TEKAX, YUCATÁN. *Tesis de pregrado].* [Universidad Autónoma de Yukatán, Yukatán.

CON-SIPA. (2018). *Análisis de Estrés de voz*. Obtenido de CON-SIPA: https://www.consipa.com/assets/vsa.pdf

Cortés, C. (2017). Herramientas Modernas en Redes Neuronales. *Universidad Autónoma de Madrid*.

Ekman, P. (2013). *Cómo detectar mentiras.* Editorial Paidós. Obtenido de https://www.academia.edu/25902701/Como_detectar_mentiras_Paul_Ekman

Gobierno de Canarias. (noviembre de 2016). *LA VOZ HUMANA*. Obtenido de Gobiernodecanarias: http://www3.gobiernodecanarias.org/medusa/ecoblog/jsancabc/files/2016/11/la-voz-humana.pdf

Gromé, M. (26 de junio de 2019). *Partes y funciones del sistema respiratorio*. Obtenido de Unprofesor: https://www.unprofesor.com/ciencias-naturales/partes-y-funciones-del-sistema-respiratorio-2280.html

Guzmán, M. (25 de marzo de 2021). *LA FUNCIÓN DE NUTRICIÓN: APARATO RESPIRATORIO*. Obtenido de Pictoeduca: https://www.pictoeduca.com/leccion/193/aparato-respiratorio/pag/1350

Hopkins, C., Benincasa, D., Ratley, R., & Griego, J. (2015). Evaluation of Voice Stress Analysis Technology. *Proceedings of the 38th Hawaii International Conference on System Sciences*, 8(5): 7695-2268. Obtenido de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.4219&rep=rep1&type=pdf

Ieda. (19 de septiembre de 2016). *Fonación*. Obtenido de Junta de Andalucia: http://agrega.juntadeandalucia.es/repositorio/19092016/71/es-an_2016091912_9102302/3_fonacin.html

Itelligent. (3 de enero de 2018). *Deep learning & Convolutional Neuronal Network: qué es y en qué consiste*. Obtenido de Itelligent: https://itelligent.es/es/deep-learning-convolutional-neuronal-network-cnn-consiste/

Kingma, D., & Ba, J. (2015). Adam: A Metrod For Atochastic Optimization. *ICLR* , 1-15.

Koerich, A. (2019). End-to-End Enviromental Sound Classification using a 1d Convolutional Neural Network.

Liji, T. (27 de febrero de 2019). *Temblores de la esclerosis múltiple*. Obtenido de News Medical: https://www.news-medical.net/health/Multiple-Sclerosis-Tremors-(Spanish).aspx

Livingstone, S., & Russo, F. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *Plos One*, 35-70. Obtenido de https://psycnet.apa.org/record/2018-23803-001

Llavina, N. (21 de enero de 2013). *Los efectos del estrés en la voz*. Obtenido de Consumer: https://www.consumer.es/salud/los-efectos-del-estres-en-la-voz.html

Lyons, R. G. (2004). *Understanding Digital Signal Processing (2nd Edition).* Nueva Jersey: Prentice Hall.

Martínez, E. (2006). DISEÑO DE BANCO DE FILTROS PARA MODELAR LA MEMBRANA BASILAR EN UNA PRÓTESIS COCLEAR. *ResearchGate*.

Martínez, G. (2013). Reconocimiento de voz basado en MFCC, SBC y espectrogramas. *Ingenius Revista de Ciencia y Tecnología*.

MathWorks. (2021). *ReluLayer*. Obtenido de MathWorks: https://www.mathworks.com/help/deeplearning/ref/nnet.cnn.layer.relulayer.html

MathWorks. (2021). *Speech Command Recognition Using Deep Learning*. Obtenido de MathWorks: https://la.mathworks.com/help/deeplearning/ug/deep-learning-speech-recognition.html;jsessionid=d26306e852a5970d2f42872b9cb0

Matich, D. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones.

Máxima, J. (2020). *Aparato Fonador*. Obtenido de Característica.co: https://www.caracteristicas.co/aparato-fonador/

McEwen, B., & Sapolsky, R. (febrero de 2010). *El estrés y su salud*. Obtenido de Hormone Health: https://www.hormone.org/pacientes-y-cuidadores/el-estres-y-su-salud

Navia, A. (29 de noviembre de 2018). *¿Cómo funcionan las Convolutional Neural Networks? Visión por Ordenador*. Obtenido de Aprendemachinelearning: https://www.aprendemachinelearning.com/como-funcionan-las-convolutional-neural-networks-vision-por-ordenador/

Nonó, J., Plaja, A., Pagés, E., Corbella, C., & Santamaria, M. (2015). *El uso profesional de la voz.* Catalunya: Departamento de Empresa y Ocupación.

Ortiz, M. (12 de diciembre de 2012). *CÓMO DETECTAR MENTIRAS*. Obtenido de COP: https://www.cop-cv.org/db/docu/1212041bero1ZfI4eaU7r7.pdf

Peña, O. (2019). ANÁLISIS DE AUDIO PARA EXTRACCIÓN DE CARACTERÍSTICAS, SEGMENTACIÓN, CLASIFICACIÓNY PREDICCIÓN. *Tesis de pregrado].* [Centro de Investigacion en Matemáticas A.C., Guanajuato. Obtenido de https://cimat.repositorioinstitucional.mx/jspui/bitstream/1008/1029/1/TE%20754.pdf

Rincón, C. (2007). Diseño, implementación y evaluación de técnicas de identificación de emociones a través de la voz. *Universidad Politécnica de Madrid*.

Rodriguez , E. (febrero de 2019). *Detectar mentiras con inteligencia artificial*. Obtenido de Investigación y Ciencia: https://www.investigacionyciencia.es/revistas/investigacion-y-ciencia/el-test-cuntico-definitivo-759/detectar-mentiras-con-inteligencia-artificial-17150

Sagasta. (2012). *Lavoz humana y su fisiología. Clasificación de las voces*. Obtenido de Preparadores de oposiciones para la Enseñanza: https://www.preparadores.eu/secundaria/Musica/Musica-Tema.pdf

Salcedo, D. (2006). Diseño de un sistema de reconocimiento del habla para controlar dispositivos eléctricos. *Universidad Católica Andrés Bello*.

Salinas, E. (2020). Digitalización de la Voz.

Sampieri, R., Fernández, C., & Baptista , P. (2010). *Metodología de la Investigación.* México: Industria Editorial Mexicana. Obtenido de https://www.uv.mx/personal/cbustamante/files/2011/06/Metodologia-de-la-Investigaci%C3%83%C2%B3n_Sampieri.pdf

Sanz, J. (20 de julio de 2018). *¿Funcionan los programas para detectar mentiras a través del análisis de la voz?* . Obtenido de Club del Lenguaje No Verbal: https://comportamientonoverbal.com/clublenguajenoverbal/funcionan-los-programas-para-detectar-mentiras-a-traves-del-analisis-de-la-voz-club-del-lenguaje-no-verbal/

Son, D. (26 de noviembre de 2010). *Producción de la Voz*. Obtenido de Zonic Studio: https://sites.google.com/site/zonicstudio/canto/produccion-de-la-voz

Tayupanta, L. (2019). Implementación de un clasificados de géneros musicales ecuatorianos mediante Deep Learning. *Tesis de Pregrado].* [Universidad de las Fuerzas Armadas ESPE, Sangolquí.

Tiwari, V. (2010). MFCC and its applications in speaker recognition. *International Journal on Emerging Technologies*, 2.

Vozalia. (5 de octubre de 2020). *Resonadores. Resonancia de la voz. Los resonadores corporales*. Obtenido de Universidad Camilo José Cela: https://www.vozalia.com/entrenamiento-de-voz/resonadores-resonancia-de-la-voz-los-resonadores-corporales/

Xianfeng, L. (2005). Voice Stress Analysis: Detection of Deception. *Tesis de pregrado].* [Universidad de Sheffield, Sheffield. Obtenido de https://dokumen.tips/documents/voice-stress-analysis.html

Yang, J. (2018). *Music Genre Classification With Neural Networks: An Examination Of Several Impactful Variables.*