ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Reconocimiento de trastornos en la voz a partir de audios asociados a la enfermedad de Parkinson

PROYECTO INTEGRADOR

Previo la obtención del Título de:

INGENIERO EN CIENCIAS COMPUTACIONALES

Presentado por:
Michael Xavier Arce Sierra
Josué Alberto Tomalá Pozo

GUAYAQUIL - ECUADOR
Año: 2022 - 2023

DEDICATORIA

Dedico este trabajo a mis padres, que me han dado su tiempo, amor y esfuerzo para ser lo que soy. A mi familia y los mentores que me ha dado la vida, por sus excelentes consejos y su guía.

Josué Alberto Tomalá Pozo

Michael Xavier Arce Sierra

AGRADECIMIENTOS

Agradezco a Dios por ser mi guía principal en este camino, y abrirme las puertas a nuevos retos y oportunidades cada vez que lo necesitaba. Gracias a mi familia por estar siempre conmigo y apoyarme en todo lo que necesitaba. Finalmente, gracias a mis amigos que hicieron que esta etapa de mi vida sea más llevadera, en especial a mi compañero de tesis Michael Arce con quien he crecido profesionalmente desde los primeros semestres de la carrera.

Josué Alberto Tomalá Pozo

Gracias a todos los que me han acompañado en este camino, en especial a mi familia y amigos, quienes son la principal motivación de todos mis objetivos. Sin olvidarme de aquellas personas que lamentablemente ya no están a mi lado. Mencionar a mi compañero Josué Tomalá con quien he trabajado desde temprano en la carrera y hemos mejorado cada vez más en los diferentes proyectos.

Michael Xavier Arce Sierra

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; Josué Alberto Tomalá Pozo y Michael Xavier Arce Sierra damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Firmado electrónicamente por:
MICHAEL XAVIER ARCE SIERRA

Michael Xavier Arce Sierra

Firmado electrónicamente por:
JOSUE ALBERTO TOMALA POZO

Josué Alberto Tomalá Pozo

EVALUADORES

LUCIA MARISOL VILLACRES FALCONI
Digitally signed by LUCIA MARISOL VILLACRES FALCONI
Date: 2023.02.14 17:56:32 -05'00'

Lucia Marisol Villacrés, Ph.D.

PROFESOR DE LA MATERIA

Firmado electrónicamente por:
COLON ENRIQUE PELAEZ JARRIN

Enrique Peláez Jarrín, Ph.D.

PROFESOR TUTOR

RESUMEN

La Enfermedad de Parkinson es la segunda enfermedad neurodegenerativa más común en el mundo, que actualmente afecta a más de 10 millones de personas. Aunque no tiene cura, existen tratamientos que logran minimizar el impacto que tiene en la vida de los pacientes y sus familias. Por esto, es esencial su detección temprana. Un síntoma característico de esta enfermedad es la debilidad de los músculos que producen la voz, lo que genera cambios en su forma de hablar. Bajo esta premisa se propone un modelo, basado en técnicas de aprendizaje autónomo y aprendizaje profundo, para detectar esos cambios como patrones asociados a la enfermedad, y determinar la probabilidad de que una persona tenga Parkinson.

Para el desarrollo del proyecto se obtuvo las grabaciones de audios de sujetos diagnosticados con Parkinson, desde 3 fuentes diferentes: 2 clínicas neurológicas en España y Guayaquil, y un grupo de personas voluntarias. Los archivos de audios pertenecen a grabaciones de personas entonando el fonema A, fonema M y leyendo una lectura corta. De estos audios se extrajeron espectrogramas de frecuencias que fueron los datos de entrenamiento de una red neuronal convolucional tipo ResNet 50 V2, la cual fue la que dio mejores resultados frente a otras arquitecturas evaluadas, con un 95% de precisión en los audios de fonemas y un 97% de precisión con las lecturas cortas. Para probar este modelo, se desarrolló una aplicación web donde se puede cargar archivos de audio de nuevas personas y obtener un diagnóstico de la enfermedad. Como resultado, se logró grabar audios de pacientes con diagnóstico de Parkinson y de sujetos sanos, pre-procesarlos y analizarlos, mediante un modelo basado en deep learning que detecta la enfermedad e integrado a una aplicación web.

Palabras Clave: Enfermedad de Parkinson, inteligencia artificial, deep learning, trastornos de voz, espectrograma.

ABSTRACT

Parkinson's disease is the second most common neurodegenerative disease in the world, currently affecting more than 10 million people. Although there is no cure, there are treatments that manage to minimize the impact it has on the lives of patients and their families. For this reason, its early detection is essential. A characteristic symptom of this disease is the weakness of the muscles that produce the voice, which generates changes in the way you speak. Under this premise, a model is proposed, based on autonomous learning and deep learning techniques, to detect these changes as patterns associated with the disease, and determine the probability that a person has Parkinson's.

For the development of the project, audio recordings of subjects diagnosed with Parkinson's were obtained from 3 different sources: 2 neurological clinics in Spain and Guayaquil, and a group of volunteers. The audio files belong to recordings of people intoning the phoneme A, phoneme M and reading a short reading. From these audios, frequency spectrograms were extracted, which were the training data of a ResNet 50 V2-type convolutional neural network, which was the one that gave the best results compared to other architectures evaluated, with 95% accuracy in the audios of phonemes and 97% accuracy with short readings. To test this model, a web application was developed where you can upload audio files of new people and get a diagnosis of the disease. As a result, it was possible to record audios from patients diagnosed with Parkinson's and from healthy subjects, pre-process and analyze them, using a model based on deep learning that detects the disease and integrated into a web application.

Keywords: Parkinson's disease, artificial intelligence, deep learning, voice disorders, spectrogram.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

La enfermedad de Parkinson (EP) es la segunda enfermedad neurodegenerativa más común luego del Alzheimer, presente en casi 9 millones de personas a nivel mundial, según la Organización Mundial de la Salud [1]. Según Parkinson's Foundation, solo en Estados Unidos de América, existen cerca de un 1 millón de personas que padecen la enfermedad, y se estima que este número aumente a 1.2 millones de pacientes en 2030 [2]. Esta enfermedad afecta directamente a la calidad de vida de los pacientes por síntomas como temblor en reposo, rigidez, demencia, depresión entre otros [3].

Actualmente se han desarrollado algunas soluciones que permiten ayudar a los doctores a diagnosticar los diferentes síntomas de la EP que suelen ser imperceptibles para el ser humano, como es la disartria, disfonía, hipofonía, entre otras [3]. En T. J. Wroge et al. [4], usaron grabaciones de fonemas de la voz de los pacientes de la EP, y a partir de biomarcadores y un modelo de árbol de decisión potenciado por gradiente se realiza un diagnóstico de la enfermedad con una precisión del 86%. En [5] los autores muestran otros métodos de inteligencia artificial (IA) para realizar el diagnóstico a través de grabaciones de audio de fonemas, entre ellos: Logistic Regression, Support Vector Machine, Multilayer Perceptrón, Boosting Regression Tree, Langragian Support Vector Machine, entre otros. Lo importante es que estos modelos fueron entrenados con características específicas de la voz del paciente como: acústica de la voz, fonación y articulación, lingüística, disfonía, características directas de la voz, frecuencias de tiempo, velocidad de habla, entre otras.

Por otro lado, otras soluciones se enfocan en identificar cambios en las expresiones faciales de los pacientes, ya que con el tiempo estas se tornan rígidas y notoriamente forzadas [6]. Sin embargo, esta solución se puede usar cuando el paciente tiene un gran avance en la enfermedad ya que es uno de los últimos síntomas en mostrarse [7].

El presente trabajo se enfoca en analizar los trastornos que sufre la voz y la forma de lectura en los pacientes con Párkinson. Este método no invasivo podría ayudar al médico a percibir estos cambios en la voz que no suelen ser notorios al oído humano.

Para esto usaremos métodos basados en el Aprendizaje Autónomo (AA), donde un algoritmo permite clasificar a un paciente con Parkinson positivo o negativo con cierta probabilidad.

Este capítulo está organizado de la siguiente manera, en la sección 1.1 se detalla el problema principal de la EP, junto con la sección 1.2 donde se justifica la importancia del desarrollo del presente proyecto y su impacto. En la sección 1.3 se presentan el objetivo general y los objetivos específicos del proyecto. Por último, en la sección 1.4 se presenta el marco teórico que explica otros aspectos importantes de la EP, el estado del arte de las soluciones y los trabajos relacionados a este proyecto.

1.1 Descripción del problema

La EP está presente aproximadamente de 0.5% a 1% en personas de 65 a 69 años, y aumentando de 1% a 3% en personas de 80 años o mayor [8]. La EP crea una pérdida parcial o total en los reflejos de la capacidad motora, la comunicación, el procesamiento del razonamiento entre otras esenciales, lo cual empeora la calidad de vida de los pacientes [7]. Por ello, se torna importante proveer mecanismos de detección temprana de la enfermedad; esta detección depende de la etapa en la que se presente la enfermedad. En [9], se especifican dos etapas de la EP: Parkinson Precoz y Avanzado. Para cada etapa existen tratamientos específicos para los síntomas motores y no motores. Es importante identificar la etapa en la que se encuentra el paciente, debido a que, si se encuentra en la etapa Avanzada, será más difícil conseguir un buen control sintomático, sin efectos adversos o complicaciones.

Una de las características identificables desde la etapa precoz de la enfermedad es que las personas con esta enfermedad sufren de alteraciones en el lenguaje, como la disfonía, hipofonía, monótono y disartria. Estas alteraciones están presentes entre el 70% al 90% de los pacientes [3]. Estas alteraciones en etapas tempranas de la enfermedad pueden llegar a ser imperceptibles para el oído humano, por lo que se usan los recursos que nos ofrece actualmente la IA para poder analizar profundamente grabaciones de audios de los pacientes. Además, la calidad de los registros de las alteraciones del lenguaje, a través de dispositivos de comunicación, como una computadora o el teléfono móvil, es alta y de bajo costo;

así como fácil de usar por cuenta propia. Por lo que, para este tipo de diagnóstico, no es necesario que el paciente este presente, haciendo el diagnóstico de la enfermedad accesible a pacientes con dificultades de movilidad.

1.2 Justificación del problema

Aunque la EP es un trastorno común, hay otros trastornos que comparten los mismos síntomas de temblores, lentitud de movimientos y rigidez [3]. Debido a esto se crean complicaciones para un buen diagnóstico y tratamiento. Según la organización mundial de la salud [1] la prevalencia de la enfermedad de Parkinson se ha duplicado en los últimos 25 años.

Aunque la enfermedad no tiene cura, existen medicinas, tratamientos y procedimientos quirúrgicos que pueden aliviar los síntomas de la enfermedad [10], [11]. Por esto su diagnóstico es tan importante que la OMS ha desarrollado un plan de acción mundial para minimizar la brecha entre las personas y los servicios de diagnóstico de enfermedades como la de Parkinson [1]. Una de las recomendaciones es la telemedicina, donde ésta propuesta de diagnóstico a través de grabaciones de la voz, puede ser de gran ayuda para el servidor de salud.

Existen métodos basados en IA y la voz que ayudan a la detección temprana de la enfermedad [4], [5]. Sin embargo, existen nuevas características como el ritmo, velocidad de lectura, vocalización, entre otros, que permiten mejorar el desempeño de los algoritmos de clasificación al momento de diagnosticar a un paciente como EP positivo o negativo.

1.3 Objetivos

1.3.1 Objetivo General

Desarrollar un modelo predictivo que permita reconocer trastornos en la voz asociados a la EP, a partir del análisis de grabaciones de la voz de pacientes diagnosticados con la enfermedad y de sujetos sanos como sujetos de control, utilizando técnicas de Aprendizaje Profundo y Aprendizaje de Máquina para ayudar a los proveedores de salud en el diagnóstico temprano de la enfermedad.

1.3.2 Objetivos Específicos

- Realizar grabaciones de pacientes con la EP y sujetos sanos de diferentes fuentes para usarlas en el entrenamiento del modelo.
- Extraer y evaluar las características auditivas más importantes para realizar el análisis y comparación de los audios.
- Identificar el mejor modelo de IA para predecir la probabilidad de que una persona padezca la EP.
- Desarrollar una aplicación web para visualizar los resultados de la evaluación de los archivos de audio y clasificarlos como Parkinson positivo o negativo según el audio ingresado, haciendo uso del modelo entrenado.

1.4 Marco teórico

1.4.1 Enfermedad de Parkinson

La EP se caracteriza por la presentación de temblor en reposo, acinesia, rigidez e inestabilidad postural [3]. Sin embargo, dependiendo de la etapa en la que se encuentra el paciente, puede presentar otros síntomas, tanto motores como no motores [9]. La EP es detectada por el médico a través de un análisis de las principales manifestaciones clínicas de la enfermedad, como las que se presentan a continuación [3]:

- Manifestaciones cardinales: Temblor de reposo, Rigidez, Acinesia, Inestabilidad postural.
- Manifestaciones no motoras: Déficit cognitivos sin demencia, Demencia, Depresión, ansiedad, Trastornos del sueño.
- Alteraciones oculares.
- Disartria.
- Disfagia.
- Anomalías olfatorias.
- Alteraciones autonómicas: Estreñimiento, Hipotensión ortostática, Alteración de la función sexual, Alteraciones urinarias.
- Sensitivas: Dolor, parestasias, calambres.
- Dermatológicas: Seborrea.

Según [3], estos síntomas aparecen cuando el 80% de las neuronas dopaminérgicas nigrales se han degenerado, debido a esto el inicio de la enfermedad es lento y progresivo. En la etapa precoz de la EP, se pueden presentar las siguientes alteraciones motoras: temblores, rigidez, acinesia, alteración de la marcha e inestabilidad postural, alteraciones oculares y alteraciones orofaríngeas; siendo las alteraciones orofaríngeas las más comunes [3].

El 90% pacientes con EP presentan trastornos en el habla. Entre estos trastornos se incluyen la acinesia faríngea, disfagia y disartria. Esta última, también llamada disartria hipocinética, se caracteriza por presentar rigidez, bradicinesia y control muscular reducido en la laringe, órganos articulatorios y otros mecanismos necesarios para realizar el habla [12]. Algunos síntomas que el paciente puede presentar con la disartria hipocinética son los siguientes [13]:

- Disminución de la intensidad vocal (hipofonía).
- Ronquea, voz áspera.
- Aumento de nasalidad.
- Habla monótona (disprosodia).
- Trastorno de velocidad del habla.
- Trastorno de fluidez.

Estos síntomas provocados por la disartria pueden ser los indicadores más tempranos para identificar la EP [14]. Por ello, diferentes investigaciones han utilizado la identificación de estos síntomas en la voz de los pacientes para diagnosticar la enfermedad.

1.4.2 Características de la voz

Un mecanismo utilizado para la obtención de características identificables en la voz es la herramienta computacional OpenSMILE [15], la cual permite obtener características estadísticas del habla. Esta herramienta permite ingresar un archivo de audio por medio de línea de comandos, extraer características y retornarlas en un archivo en formato CSV [15].

1.4.3 Soluciones implementadas

Utilizando estas características se han realizado soluciones utilizando técnicas de IA. En [16], se realizó un análisis comparativo entre diferentes modelos basados en IA entre los que se mencionan: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), entre otros; donde el modelo SVM es el que obtiene mejor precisión. Así mismo pasa en [14], [17], [18] donde SVM es de los mejores clasificadores.

SVM es un algoritmo supervisado que permite realizar clasificación o regresión, utiliza optimización convexa, teorías de aprendizaje estadísticos y máquinas kernel. Este algoritmo fue presentado por primera vez en [19], y ha ido evolucionando hasta las implementaciones que actualmente existen [20].

En [17] también usan Principal Component Analysis (PCA) [21], este es un método para poder reducir la dimensionalidad de los datos que poseen muchas características, aumentando la interpretabilidad y minimizando la pérdida de información. Este procedimiento permite redimensionar las características de manera que la información que ingresará al modelo es la que mejor representa a toda la lista de características inicialmente obtenidas, en nuestro caso las características de la voz del sujeto [17].

1.4.4 Espectrogramas de la voz

Un espectrograma es una representación gráfica del espectro de frecuencias de, en este caso, la voz. En él se pueden observar las altas frecuencias, potencia o modulaciones de amplitud, las cuales no se pueden apreciar con el oído humano [22].

Utilizando espectrogramas, es posible detectar diferencias en la voz entre sujetos sanos y pacientes de EP. Con ellos, se puede observar que la frecuencia media, mínima y máxima de la voz de los pacientes de EP es menor que los sujetos sanos, como también que la desviación estándar de la frecuencia fundamental se ve reducida en los pacientes [23], como se puede observar en la Figura 1.

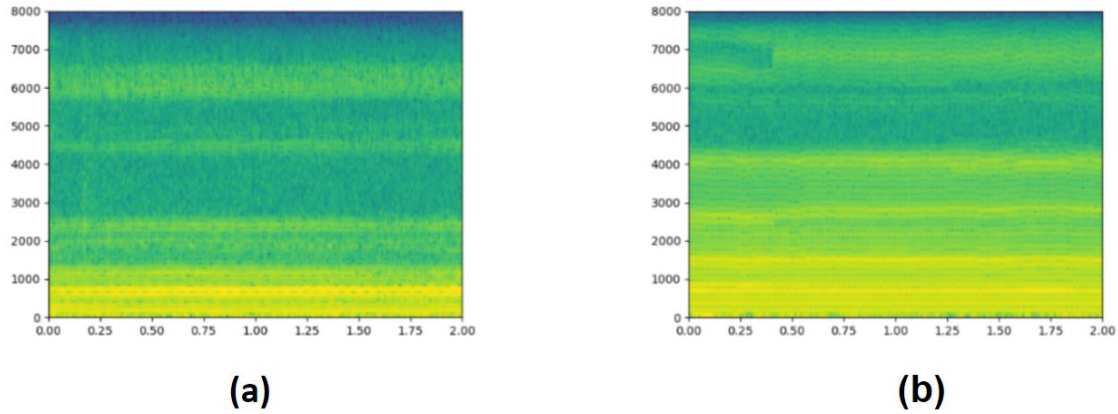

Figura 1: Espectrogramas de frecuencia de (a) un paciente y (b) un sujeto sano. (Autoría propia)

CAPÍTULO 2

En este capítulo se describe el procedimiento para la obtención de audios y la selección de características de la voz para el entrenamiento del modelo. También se describen los modelos de aprendizaje autónomo utilizados. Finalmente, se presenta la implementación de la página web, junto con las herramientas utilizadas.

2. METODOLOGÍA

La metodología empleada está dividida en 5 etapas secuenciales, es decir, por ejemplo, las actividades realizadas en la etapa 2 dependen de la etapa 1, como se puede observar en la Figura 2. Las etapas son: Obtención de audios, preprocesamiento de los audios, extracción de características de los audios, selección del modelo basado en IA a utilizar y, por último, implementación de la página web.

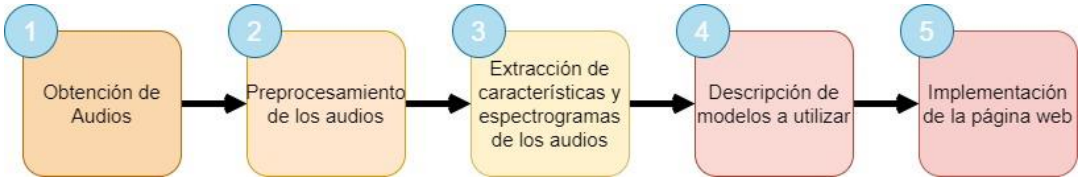

Figura 2: Etapas de la metodología aplicada (Autoría propia)

2.1 Obtención de audios

Se obtuvo 3 conjuntos de datos, de los cuales: el primer conjunto denominado A1 proviene de grabaciones realizadas en el Centro Integral en Neurociencias – CINAC de HM Hospitales de la Universidad San Pablo. El segundo conjunto proviene de pacientes provistos por el Dr. Juan Carlos Solá durante sus procesos de consulta en el hospital Clínica Kennedy de la ciudad de Guayaquil; en ellas se grabó el audio y video de pacientes diagnosticados con EP, mediante una aplicación realizada en Python, donde el paciente debía entonar fonemas, así como leer un texto corto, como se muestra en el ANEXO 1.

Este conjunto se denominó A2. Por último, se realizaron grabaciones a personas sanas cercanas a los autores de este documento, que tenían edades entre 50 y 75 años (edades similares a la de las personas del conjunto A1 y A2). A este último conjunto se lo denominó A3.

2.2 Descripción de los audios

Los fonemas obtenidos en el conjunto A2 y A3 son entonaciones prolongadas de las letras “a” y “m”; mientras que en los conjuntos A1 solo de la letra “a”. Se utilizaron estos fonemas debido a que, como se indica en [6], presentaron mejores resultados de precisión a la hora de identificar a personas con EP. También, en los conjuntos A1, A2 y A3, se realizó la grabación de un texto leído por los pacientes. Este texto era el siguiente: “Platero es pequeño, peludo, suave; tan blando por fuera, que se diría todo de algodón, que no lleva huesos. Sólo los espejos de azabache de sus ojos son duros cual dos escarabajos de cristal negro.” Este fue el texto utilizado por el CINAC, por ello, para las grabaciones realizadas en la ciudad de Guayaquil, se utilizó el mismo texto.

2.3 Preprocesamiento de audios

Primero, se editaron manualmente los audios grabados de los conjuntos A2 y A3. En los audios de este conjunto, se encontraban la entonación de los fonemas “a” y “m”, y la lectura del texto. Por ello, se recortaron los audios removiendo los espacios en que el paciente no habla. Dejando como resultado tres extractos por cada audio de A3. El resto de los audios del conjunto A1 ya se encontraban recortados, de tal manera que el contenido de los archivos era solamente el fonema y el texto leído.
Luego, se procedió a realizar un aumento de datos de estos audios, tanto de fonemas y textos. Utilizando el lenguaje Python, se extrajeron segmentos de dos segundos de los audios, dejando una diferencia de un segundo entre los segmentos.

2.4 Extracción de características

Luego de preprocesar los audios, se procedió a extraer las características de la voz que servirían para alimentar al modelo de inteligencia artificial. Las características no son las mismas para los fonemas y los textos. Las características extraídas para el análisis de fonemas se realizó con OpenSmile [15], al igual que con la extracción de características para el análisis de textos. Las características obtenidas, tanto para fonemas como textos, tal como lo propusieron en [4] fueron 6373. Una vez obtenidas, se utilizó el análisis PCA para reducir el número de características que se utilizarían para entrenar el modelo de inteligencia artificial. Utilizando PCA, se seleccionaron un total de 100 componentes, debido a que se deseaba mantener el

95% de la información. Estos componentes fueron utilizados para entrenar los modelos.

2.5 Representación de audios en espectrogramas de frecuencia

También, con los audios pre-procesados, se extrajo la representación en espectrogramas de los segmentos de audio. Utilizando las librerías, de Python, Librosa y Matplotlib se crearon las imágenes en formato PNG de 180x320 pixeles de tamaño, a partir de los espectrogramas de frecuencia extraídos desde los archivos de audio de dos segundos, donde el rango de frecuencias iba desde 0 hasta 8000 Hz, debido a que las pruebas realizadas con los modelos de aprendizaje profundo fueron con ventanas de frecuencia de: 0 a 2000, 1000 a 4000, 0 a 4000 y 0 a 8000 Hz.

2.6 Evaluación y selección de arquitecturas

En esta etapa realizamos pruebas con los modelos de predicción escogidos que usarán como entrada las características extraídas por los audios y las imágenes de espectrogramas. Como resultado obtuvimos la probabilidad de clasificación si pertenece al conjunto de sujetos con la EP o a un sujeto sano. Para la selección del mejor modelo se compararon las métricas de desempeño de los modelos: una matriz de confusión la cual nos permite observar el porcentaje de los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos; con lo que pudimos identificar qué tan acertada es la clasificación que realiza el modelo. Así mismo, se usó la curva ROC, la cual es una la representación de la razón o proporción de verdaderos positivos frente a la razón o proporción de falsos positivos, tal como se describe en la sección 2.6.8.

Las características extraídas de OpenSMILE son de tipo tabular, de manera que se usaron los modelos de clasificación basados en técnicas de aprendizaje autónomo poco profundos, o tipo shallow, tales como Support Vector Machines, XGBoost, Multi Layer Perceptron, Random Forest o Logistic Regression, los cuales se describen en las siguientes secciones.

Mientras que, para evaluar los espectrogramas obtenidos, se utilizaron los modelos de aprendizaje profundo, o convolucionales, tales como ResNet 50, ResNet 50 v2, y ConvNet, descritas en las siguientes secciones.

2.6.1 Support Vector Machine (SVM)

Este es un algoritmo usualmente usado para regresión y clasificación. El objetivo de este algoritmo es lograr identificar un hiperplano en el espacio de n-dimensiones (donde n es la cantidad de características) a través del cual se pueda identificar claramente el borde de decisión de los datos, como se puede ver en la **Figura 3**.

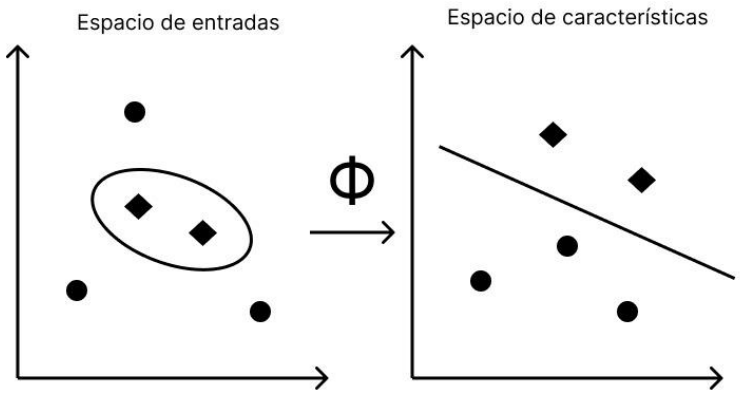

Figura 3: La idea de SVM: mapear los datos de entrenamiento de forma no lineal en un espacio de características de mayor dimensión a través de Φ, y construya un hiperplano separador con máximo margen allí. Adaptado de [24].

La principal ventaja de este algoritmo frente a otros es que nos permite trabajar con datos de gran dimensionalidad. En este caso extraemos hasta 100 características de los audios, los cuales son usados por el algoritmo para buscar el mejor hiperplano que permita realizar la clasificación.

2.6.2 XGBoost

Es un algoritmo de aprendizaje supervisado basado en arboles de decisión potenciados por gradiente. Su principal ventaja se da por la posibilidad de paralelizar tareas y aumentar su rendimiento, haciendo uso de computación paralela e incluso computación distribuida [25]. También es un algoritmo de

ensamble, lo cual significa que combina múltiples algoritmos de AM para poder obtener el mejor modelo.

Este algoritmo consiste en crear diferentes arboles de decisión, o modelos débiles que, a través del potenciado del gradiente para minimizar el error, permite ensamblarse al final del proceso para generar un modelo robusto de clasificación o regresión (Ver Figura 4).

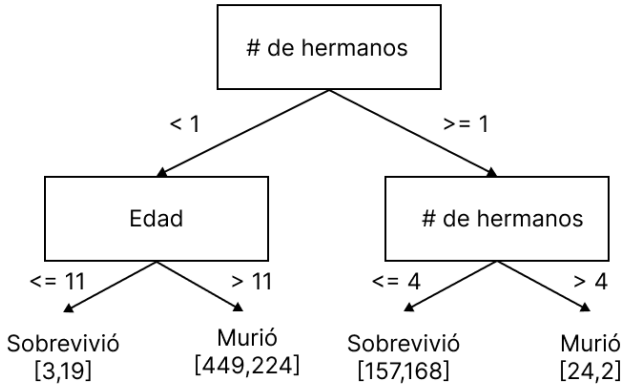

Figura 4: Un ejemplo de un árbol de decisiones que predice si es probable que un pasajero del Titanic sobreviva en función de sus atributos. Adaptado de [26]

2.6.3 Multi Layer Perceptron (MLP)

Es el tipo de arquitectura de red neuronal más básica y fundamental comparadas con otras, como las redes convolucionales y recurrentes. En esta arquitectura los perceptrones (neuronas) son apiladas en múltiples capas, cada nodo es conectado a todos los otros nodos de la siguiente capa. No existe conexión entre los nodos de una misma capa [26].

A pesar de ser una arquitectura básica de red neuronal, entrega muy buenos resultados al momento de realizar la clasificación de características (Ver **Figura 5**).

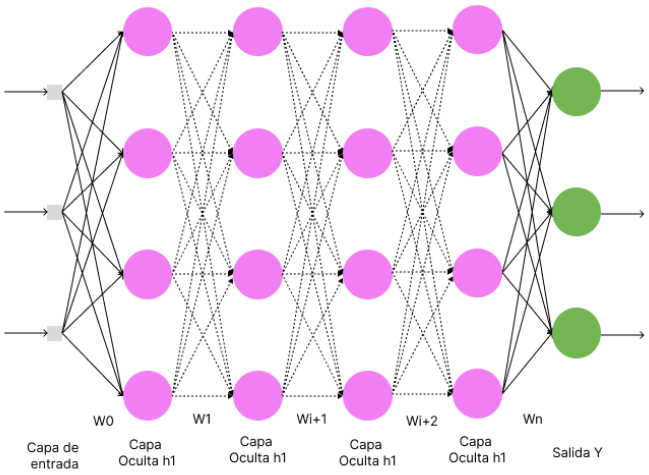

Figura 5: Un ejemplo de arquitectura de MLP. Adaptado de [26]

2.6.4 Random Forest

Este algoritmo se encarga de desarrollar un gran número de árboles de decisión que operan como un ensamble. Cada árbol retorna una predicción de clases y la clase con más votos se convierte en nuestro modelo de predicción. Este algoritmo se basa en el concepto fundamental de la sabiduría de las masas. Es decir, los errores de algunos árboles son corregidos por los otros árboles que están correctos [27]. Un ejemplo lo podemos ver en la Figura 6.

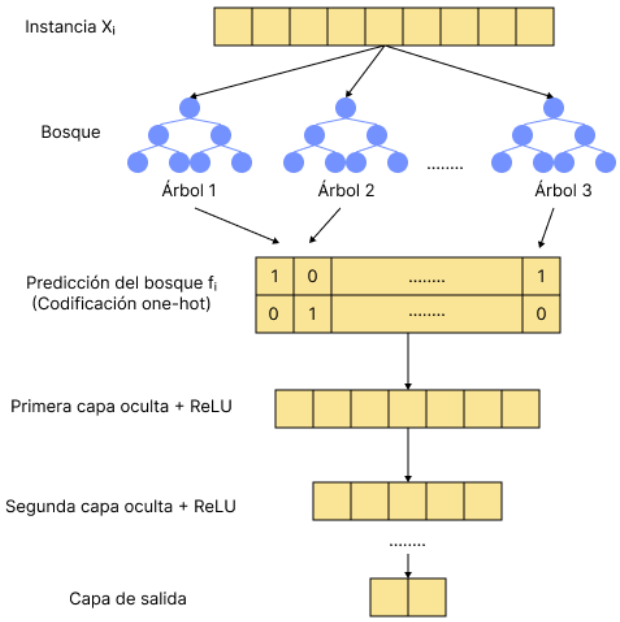

Figura 6: Representación de arquitectura de Random Forest. Adaptado de [28]

2.6.5 Regresión logística

Este algoritmo permite entrenar un modelo que predice la presencia o ausencia de características o resultado, según los valores de un conjunto de predictores. Es decir, en base a las características indicadas, el modelo retorna un indicador que determina la probabilidad que un evento exista [28]. (Ver **Figura 7**)

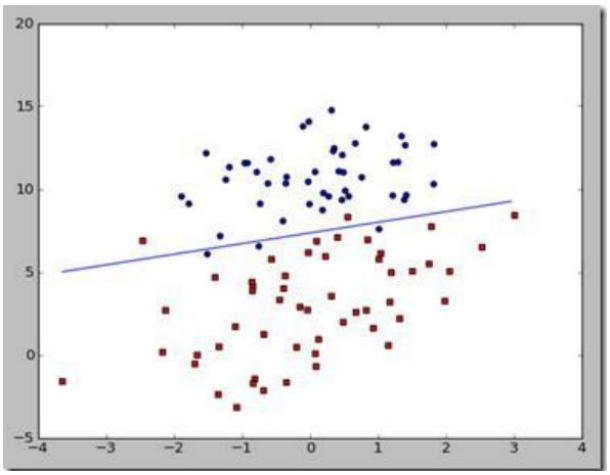

Figura 7: Representación de una regresión logística. (Fuente: [28]**)**

2.6.6 ConvNet (Red convolucional)

Es un algoritmo de aprendizaje autónomo profundo, el cual puede tomar una imagen y asignar importancias (pesos y sesgos) a varios aspectos o

características de la imagen, así diferenciando una imagen de otra. Tal como lo muestra O'Shea et al. en [29], la imagen es representada en una matriz de 3 dimensiones: Altura, anchura y el color.

Esta red está compuesta por varios bloques de procesamiento, cada una compuesta por 2 capas: una de convolución y otra de reducción llamada pooling. La capa de convolución se encarga de extraer información importante de la sección de imagen ingresada. Mientras que la capa de pooling se encarga de abstraer esta nueva información y generar una nueva matriz de información que se usará para el siguiente bloque de convolución (Ver Figura 8).

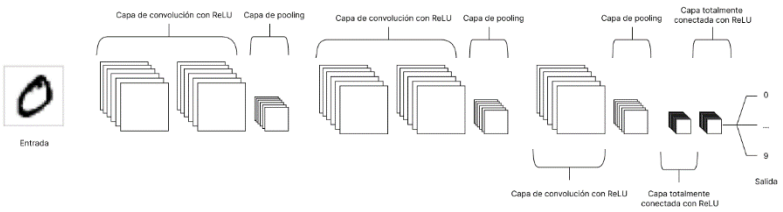

Figura 8: Red convolucional con 3 bloques de convolución. Adaptado de [29]

2.6.7 Resnet 50 V1 y V2.

Las ConvNet pueden ser efectivas usando pocos bloques de convolución. Sin embargo, cuando estos bloques aumentan, la precisión del modelo empieza a verse afectado debido al problema del desvanecimiento del gradiente. Este problema se genera debido a que, al hacer la propagación del error hacia atrás, el gradiente empieza a disminuir mucho llegando a 0, es decir la red deja de aprender efectivamente.

Para resolver este problema se desarrolló una solución con las ResNet como se presenta He et al. en [30]. Esta red genera ciertos atajos entre bloques de convolución, donde usa residuos de anteriores bloques de convolución para realizar el aprendizaje. Esto permite aumentar la cantidad de bloques a 50, 80 e incluso más de 100. Esta cantidad de bloques dependerá de la complejidad de la información que ingresamos a la red. En nuestro caso al hablarse de

espectrogramas, una red con 50 bloques de convolución pudo ser suficiente. (Ver Figura 9)

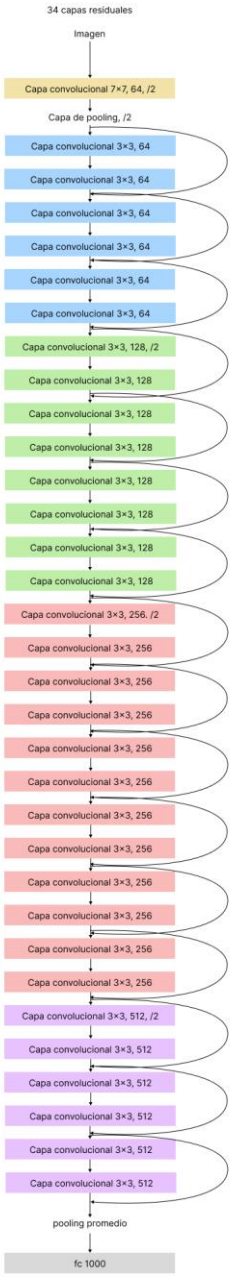

Figura 9: Ejemplo de red convolucional basada en residuos con 34 bloques de convolución. Adaptado de [30]

2.6.8 Métricas de Evaluación

Para evaluar los modelos antes expuestos se usaron las siguientes métricas:

Matriz de confusión: Esta es una representación matricial de los resultados de cada una de las predicciones, la cual se usa para describir el rendimiento del modelo de clasificación. Como se muestra en [31], esta matriz de especifica la cantidad o porcentaje de los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Para obtener la exactitud del modelo, se evalúan los resultados usando la **Ecuación 1**.

$$\boldsymbol{Exactitud = \frac{VP + VN}{VP + FP + FN + VN}}$$

Ecuación 1: Fórmula de la exactitud

Curva de características operativas del receptor (ROC): En este método se traza la tasa de verdaderos positivos (sensibilidad), calculada con la Ecuación 2, frente a la tasa de falsos positivos, calculada con la Ecuación 3. Esta traza se la conoce como curva ROC y nos permite visualizar el equilibrio entre los verdaderos positivos y falsos positivos, tal como se usa en [32]. El espacio ROC está definido por los falsos positivos como eje x y los verdaderos positivos en el eje y, y representa los intercambios entre estos.

$$\boldsymbol{TVP = \frac{VP}{VP + FN}}$$

Ecuación 2: Fórmula de la taza de verdaderos positivos, también llamada sensibilidad o recall.

$$\boldsymbol{TFP = \frac{FP}{VN + FP}}$$

Ecuación 3: Fórmula de la taza de falsos positivos.

En las ecuaciones cada variable representa:

- VP: Verdadero positivo.
- VN: Verdadero negativo.
- FP: Falso positivo.
- FN: Falso negativo.

- TVP: Tasa de verdaderos positivos
- TFP: Tasa de falsos positivos

2.7 Desarrollo de prototipo

En esta sección, se describe el funcionamiento del producto final, los actores del sistema, la arquitectura de la solución, y un prototipo.

2.7.1 Actores del sistema

Los actores del sistema fueron definidos a través de reuniones virtuales con el cliente, donde se le presentó un primer boceto de la solución, y, por medio de sus comentarios, se identificaron los siguientes roles:

2.7.1.1 Actor Especialista

Este rol está dirigido a los médicos que usarán el sistema para realizar el diagnóstico de la enfermedad. Este sistema cuenta con las siguientes funcionalidades:

- Cargar archivos de audios del sujeto a diagnosticar.
- Ejecutar herramienta para recortar los fragmentos de audio.
- Visualizar los resultados de diagnóstico y las características extraídas de la grabación de audio.
- Ingresar observaciones a los resultados del diagnóstico entregado por el sistema.
- Mostrar los resultados históricos de los antiguos diagnósticos realizados por el especialista.

2.7.1.2 Actor administrador

Este rol está dirigido al administrador del sistema, el cual se encarga de realizar el monitoreo de los diagnósticos y el correcto funcionamiento del modelo de diagnóstico basado en inteligencia artificial.

- Crear usuarios con rol especialista.
- Visualizar los históricos de los diagnósticos realizados con sus respectivos audios y resultados.
- Revisar las observaciones realizadas por los usuarios especialistas.

2.7.2 Arquitectura del prototipo

2.7.2.1 Tecnologías

La plataforma web está formada por un frontend y un backend que es administrado por la misma tecnología:

- **Django:** Es un framework que permite administrar y servir los requerimientos tanto para el frontend (vista del usuario), como el backend (consultas, modelo y base de datos). Se eligió este framework porque nos permite realizar un rápido desarrollo a diferencia de otras herramientas.
- **Sqlite:** Es un sistema de gestión de bases de datos relacional, la cual se caracteriza por almacenar información de una manera sencilla, eficaz, potente y usando pocos recursos; haciéndola una base de datos liviana.

2.7.2.2 Diagramas de la arquitectura

En esta sección se muestran los requerimientos funcionales de la plataforma web.

Los usuarios están formados por dos tipos: Especialista y administrador (ver Figura 10). El especialista es el usuario principal del sistema, quien tiene la opción de realizar el diagnóstico de la enfermedad en los sujetos, a través de las grabaciones de audios. Por otro lado, el administrador es el encargado de monitorear los diferentes diagnósticos, además de revisar las observaciones realizadas por los especialistas.

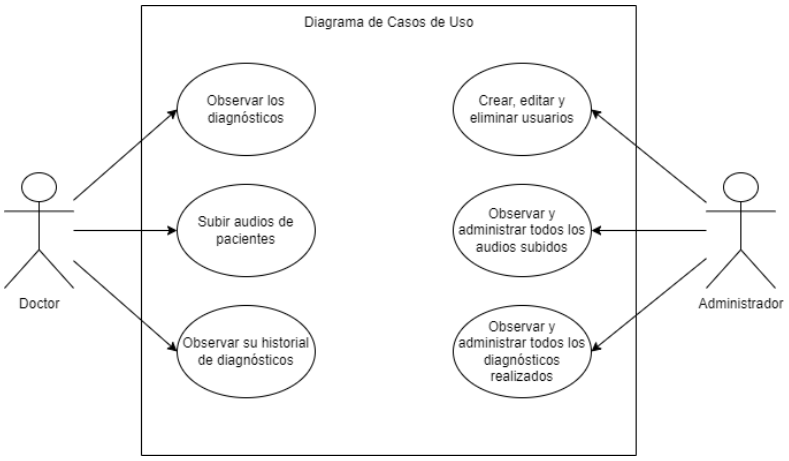

Figura 10: Diagrama de casos de uso del producto final (Autoría propia)

El sistema contiene las siguientes clases (ver **¡Error! No se encuentra el origen de la referencia.**1):

- **Usuario:** Contiene las características del sujeto que interactúa con el sistema. Además, se especifica el rol que tiene en el sistema (especialista o administrador).
- **Audio:** Contiene la información del audio y el directorio donde se encuentra ubicado el audio.
- **Diagnóstico:** Especifica la información del diagnóstico, guarda los resultados de las características, junto con la información de resultado dada por el modelo de IA.

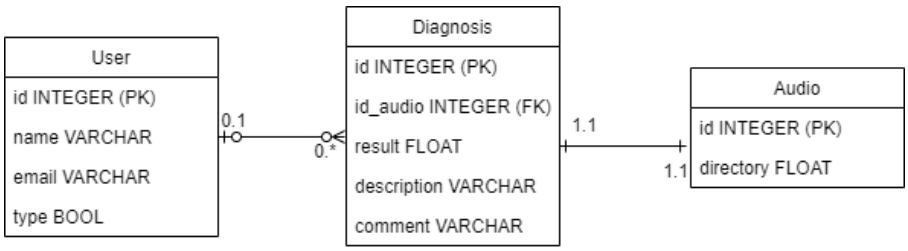

Figura 11: Diagrama Entidad-Relación (Autoría propia)

2.7.3 Prototipo Final

En esta sección se presenta el diagrama de funcionamiento del prototipo, el cual fue desplegado en la plataforma en la nube Azure que nos ofrece hosting en la nube de manera gratuito por la licencia de estudiante.

Parkinson Audio App tiene dos tipos de usuario, un rol de médico y uno de administrador. El rol médico permite cargar o grabar el audio de una persona, recortarlo para escoger el mejor segmento de audio para el análisis. Luego la aplicación genera el espectrograma el cual es usado como ingreso para ResNet. Una vez analizado el espectrograma, la aplicación muestra el resultado, mostrando que tan probable es que la persona del audio sea diagnosticada con Parkinson positivo. En esta misma pantalla el médico podrá indicar si está de acuerdo o no con el diagnóstico realizado por el modelo.

Por otro lado, si médico desea revisar los diagnósticos pasados, dentro de la sección Historial, se enlistarán todos los diagnósticos pasados.

Para el administrador, se usó el administrador de Django, el cual permite realizar cambios y acciones a todas las entidades del sistema. Así mismo se pueden realizar cambios en caso de ser necesario, y visualizar la dirección donde se encuentra el audio almacenado en el servidor donde se aloja la aplicación, que probablemente se use para futuros proyectos. Las pantallas principales se pueden observar en Figura 12, Figura 13 y Figura 14.

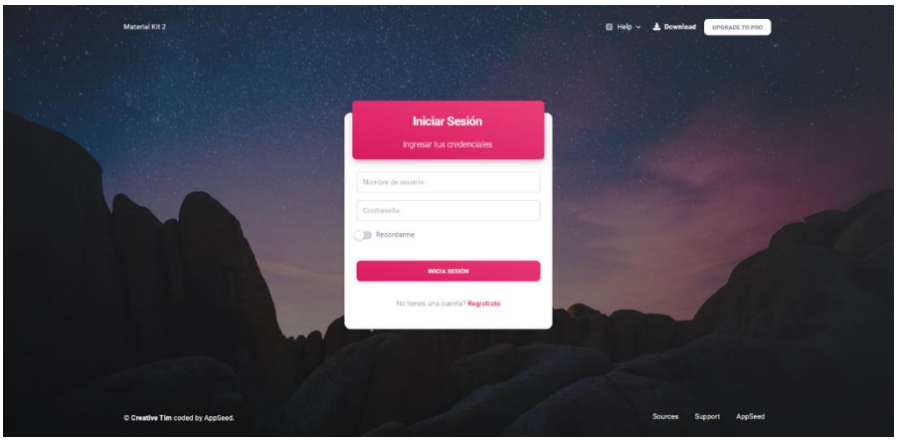

Figura 12: Pantalla de Inicio de Sesión (Autoría propia)

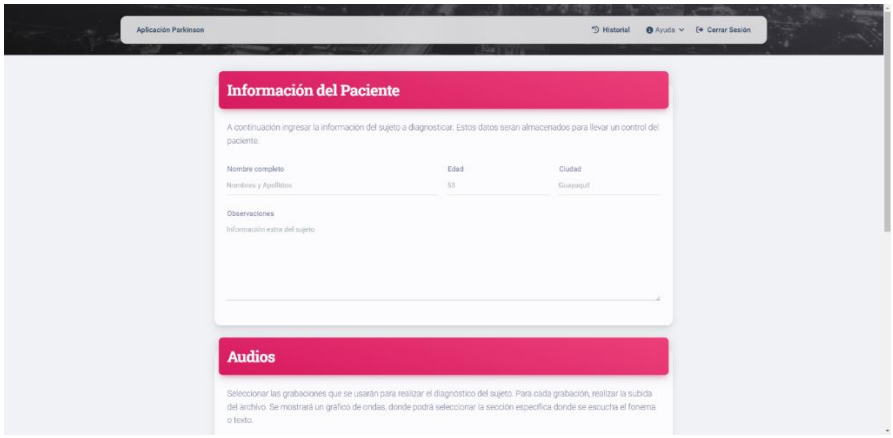

Figura 13: Formulario de ingreso de datos del sujeto (Autoría propia)

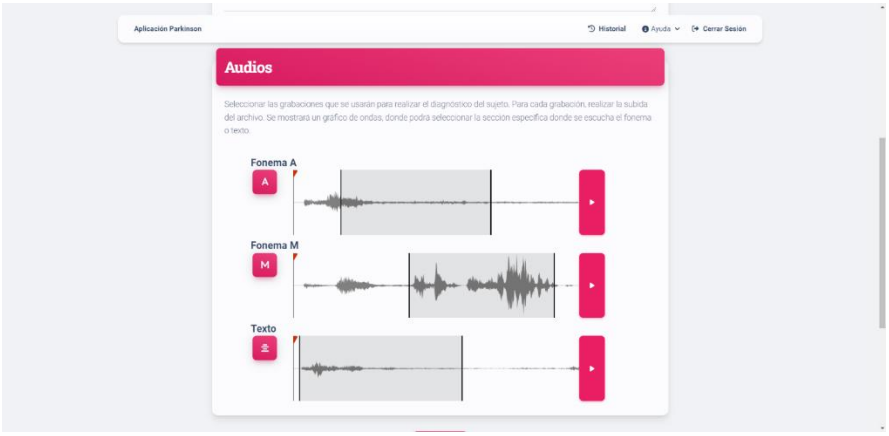

Figura 14: Ingreso y segmentación de audios (Autoría propia)

CAPÍTULO 3

3. ANÁLISIS Y RESULTADOS

En este capítulo se muestra los resultados obtenidos del entrenamiento de los modelos de inteligencia artificial mencionado en la Sección 2, y su respectiva comparación.

3.1 Experimentos desarrollados

3.1.1 Descripción del entorno de desarrollo

Se desarrollaron múltiples experimentos utilizando las librerías de Python: Keras, Tensorfow, librosa, y la herramienta OpenSmile, mencionada en la sección 2.4. Estos experimentos fueron desarrollados en una máquina local con potencia gráfica (GPU), específicamente con la tarjeta gráfica dedicada NVIDIA RTX 2060. Se hizo uso de los tres conjuntos de datos mencionados en la sección 2.2. En total, se utilizaron 186 audios de fonemas y 128 audios de textos, de los cuales se realizó un aumento de datos. De este aumento de datos, se obtuvieron 455 audios de fonemas y 405 de texto. De estos audios, se extrajeron datos para el entrenamiento de los modelos basados en aprendizaje de máquina y para el modelo de análisis de espectrogramas.

3.1.2 Datos de entrenamiento

Para los modelos basados en aprendizaje autónomo, se utilizaron tres conjuntos de datos, los cuales fueron extraídos de los tres conjuntos iniciales (ver Tabla 1)

Tabla 1: Características extraídas para modelos basados en aprendizaje de maquina

Conjunto	Tipo de datos
Todas las características	Un total de 6374 características obtenidas de OpenSMILE. Etiquetado como ALL.

Características de PCA	Un total de 100 componentes principales obtenidos por medio del uso de PCA para los audios de fonemas (etiquetado como PCA–f), y 74 componentes principales obtenidos del uso de PCA para audios de texto (etiquetado PCA–txt).

Mientras que para el análisis de espectrogramas se obtuvieron espectrogramas de potencia en distintas frecuencias, los cuales se utilizaron como imágenes de entrada para los modelos convolucionales. Se utilizó un total de 1720 imágenes, de las cuales, 910 eran espectrogramas de audios de fonemas (etiquetado como f) y 810 espectrogramas de audios de texto (etiquetado como txt).

3.1.3 Parámetros de los experimentos.

Para los experimentos realizados con los modelos entrenados con las características extraídas de los archivos de audio, se usaron los 5 modelos tipo shallow mencionados en la sección 2.6.

Para el análisis de espectrogramas se usaron 4 modelos convolucionales mencionados en la sección 2.6, con 25, 35 y 50 épocas, así mismo usando un rango de frecuencias entre: 0-2000, 0-4000, 0-8000 y 1000-4000 Hz.

3.2 Resultados de la comparación de modelos tipo shallow

De los modelos de aprendizaje autónomo explorados se obtuvieron los siguientes resultados: (ver Tabla 2)

Tabla 2: Resultados obtenidos en modelos tipo shallow

Modelo	Precisión	Conunto	Classes
SVM	0,998	PCA–f	FP, FS
XGBoost	0,995	PCA–f	FP, FS
MLP	0,989	PCA–f	FP, FS
RandomForest	0,984	PCA–f	FP, FS

LogisticRegression	0,998	PCA–txt	TP, TS
SVM	1	PCA–txt	TP, TS
XGBoost	0,998	PCA–txt	TP, TS
MLP	0,993	PCA–txt	TP, TS
RandomForest	0,985	PCA–txt	TP, TS
LogisticRegression	1	PCA–txt	TP, TS
SVM	1	ALL–f	FP, FS
XGBoost	1	ALL–f	FP, FS
MLP	1	ALL–f	FP, FS
RandomForest	1	ALL–f	FP, FS
LogisticRegression	1	ALL–f	FP, FS
SVM	1	ALL–txt	TP, TS
XGBoost	1	ALL–txt	TP, TS
MLP	1	ALL–txt	TP, TS
RandomForest	1	ALL–txt	TP, TS
LogisticRegression	1	ALL–txt	TP, TS

En base a la información de la tabla, podemos notar que los modelos entrenados con los datos obtenidos luego de aplicar PCA, presentan una precisión bastante alta. Sin embargo, los modelos entrenados con todas las 6374 características obtenidas con OpenSMILE [15], inducen a un overfitting del modelo, es decir que cuando el modelo ingrese a pruebas reales, probablemente produce resultados erróneos.

3.3 Resultados de comparación de modelos convolucionales

Para el análisis de espectrogramas se realizó un total de tres pruebas estructuradas de la siguiente manera:

- Pruebas de las diferentes arquitecturas convolucionales
- Pruebas de diferentes ventanas de tiempo para las gráficas del espectrograma.
- Pruebas de diferentes ventanas de frecuencias para las gráficas del espectrograma.
- Pruebas de la cantidad de épocas que se entrena el modelo.

3.3.1 Pruebas de arquitecturas

Para las pruebas de las diferentes arquitecturas se usó las librerías Keras y Tensorflow. Las arquitecturas seleccionadas, tal como se revisaron en la sección 2, fueron:

- ResNet 50
- ResNet 50 V2
- CNN

Tabla 3: Prueba de las diferentes arquitecturas convolucionales

Arquitectura	Épocas	Ventana espectral	Ventana temporal	Accuracy	Loss	Datos
RESNET 50	25	0-8000	6s	0.84	0.47	Fonemas A y M
RESTNET 50 V2				0.94	0.18	
CNN				0.82	1.04	
RESNET_50				0.77	2.77	Texto
RESTNET 50 V2				0.95	0.11	
CNN				0.89	0.40	

Como vemos en la Tabla 3, las mejores redes convolucionales fueron: ResNet50 v2 en primer lugar, seguido de la CNN y en tercer lugar la ResNet 50.

Dados estos resultados, las siguientes pruebas se realizaron con las arquitecturas de redes que mejor resultado dieron, en este caso la ResNet 50 v2 y la CNN.

3.3.2 Pruebas de ventanas temporales

En estas pruebas se evaluó cual es la mejor ventana temporal de tiempo que se debe usar con los audios de los pacientes. Esto permitió identificar cual será la ventana de tiempo para el entrenamiento y la predicción del modelo. Se probaron las siguientes ventanas temporales: 2, 4 y 6 segundos.

Tabla 4: Pruebas de ventanas temporales para análisis de espectrogramas

Arquitectura	Épocas	Ventana espectral	Ventana Tiempo (s)	Accuracy	Loss	Datos
RESTNET 50 V2	25	0-8000	2s	1.00	0.00	Fonemas A y M
RESTNET 50 V2			4s	0.55	3.26	
RESTNET 50 V2			6s	0.65	1.04	
RESTNET 50 V2			2s	0.94	0.18	Texto
RESTNET 50 V2			4s	0.98	0.09	
RESTNET 50 V2			6s	0.53	3.75	
CNN			2s	0.95	0.08	Fonemas A y M
CNN			4s	0.85	0.33	
CNN			6s	0.64	1.54	
CNN			2s	0.90	0.40	Texto
CNN			4s	0.87	0.30	
CNN			6s	0.58	0.97	

Como se muestra en la Tabla 4, la ventana temporal que dio mejor resultado fue de 2s. Por esto, los audios a analizar serán cortados a esos tiempos para ser entrenados. Debido a que la ventana temporal que da mejor resultados es de 2s y los audios tienen una duración mayor, de hasta 7 segundos, se usarán los audios para realizar el aumento de datos como se muestra en la sección 2.3. Con eso logramos mejorar la cantidad de audios usados para entrenar los datos.

3.3.3 Pruebas de ventana de frecuencias

En esta sección realizamos el análisis de las pruebas de la ventana de frecuencias del espectrograma. Con estas pruebas identificamos cual es el rango de frecuencias donde se muestra la diferencia más importante entre los sujetos que padecen la enfermedad, y los que no, como se menciona en la sección 1.4.4. Para estas pruebas usamos las siguientes ventanas de frecuencias: 0-2000, 0-4000, 0-8000 y 1000-4000 Hz.

Tabla 5: Pruebas de ventanas de frecuencia en análisis de espectrogramas

Arquitectura	Épocas	Ventana espectral	Ventana Tiempo (s)	Accuracy	Loss	Datos
RESTNET 50 V2	25	0-2000	2s	0.98	0.01	Fonema A y M
RESTNET 50 V2		0-4000		0.92	0.24	
RESTNET 50 V2		0-8000		0.94	0.18	
RESTNET 50 V2		1000-4000		0.99	0.05	
RESTNET 50 V2		0-2000		0.98	0.08	Texto
RESTNET 50 V2		0-4000		0.92	0.01	
RESTNET 50 V2		0-8000		0.97	0.06	
RESTNET 50 V2		1000-4000		0.56	3.95	

Como podemos notar en la Tabla 5, la mejor ventana temporal que se muestra en las diferentes arquitecturas y datos es la de 0-8000 Hz. Esto se debe a que es donde se muestra una gran diferencia entre los espectrogramas de los diferentes sujetos. Un ejemplo se muestra en la **¡Error! No se encuentra el origen de la referencia.**.

3.3.4 Pruebas de épocas

Para la prueba de épocas se usó solo la arquitectura con los mejores resultados, la cual es la ResNet 50 v2. Así mismo se usó la mejor configuración encontrada anteriormente, la cual es: Ventana temporal de 2s y ventana espectral de 0-8000 Hz. Probamos las siguientes épocas: 15, 25, 30, 35 y 50 épocas.

Tabla 6: Pruebas de épocas en análisis de espectrogramas

Arquitectura	Épocas	Ventana espectral	Ventana Tiempo (s)	Accuracy Train	Loss	Accuracy validation	Datos
RESTNET 50 V2	15	0-8000	2s	0.57	20.64	0.50	Fonema A y M
RESTNET 50 V2	25			0.98	0.01	1.0	
RESTNET 50 V2	30			0.97	0.07	0.99	
RESTNET 50 V2	35			1.00	0.00	0.99	
RESTNET 50 V2	50			1.00	0.00	1.0	
RESTNET 50 V2	15			0.54	11.04	0.50	Texto
RESTNET 50 V2	25			0.92	0.43	0.85	
RESTNET 50 V2	30			0.99	0.00	1.00	
RESTNET 50 V2	35			0.99	0.02	1.00	
RESTNET 50 V2	50			1.00	0.00	1.00	

Como notamos en la Tabla 6, la mejor cantidad de épocas para entrenar el modelo es de 25 épocas con esta cantidad de datos. No usamos las de 35 y 50 épocas debido a que podríamos inducir al modelo a sobre entrenamiento.

3.3.5 Evaluaciones Finales

Para evaluar el modelo final se utilizó la matriz de confusión y la curva ROC para visualizar que tan exacto es el modelo de clasificación, tal como se describió en la sección 2.6.8. Los datos utilizados para el entrenamiento fueron: 455 audios de fonemas de pacientes y 455 de sujetos sanos, manteniendo 50 audios para validación por cada uno, y 405 audios de texto de pacientes y 405 de sujetos sanos, manteniendo 40 audios para validación por cada uno.

Este modelo fue entrenado usando los siguientes parámetros:

- Arquitectura: ResNet 50 V2
- Ventana Temporal: 2s
- Ventana Espectral: 0-8000 Hz
- Épocas: 25

3.3.5.1 Validaciones de Fonemas

Para la validación del modelo, se usaron 50 audios elegidos de manera aleatoria. Estos audios no se usaron para entrenar los modelos, con lo que se obtiene las siguientes gráficas.

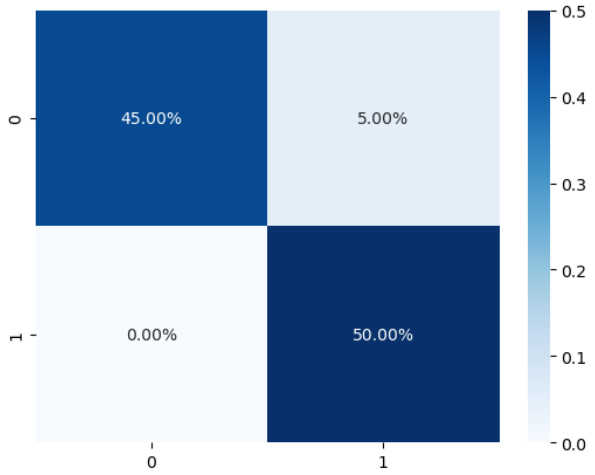

Figura 15: Matriz de confusión de validación de fonemas (Autoría propia)

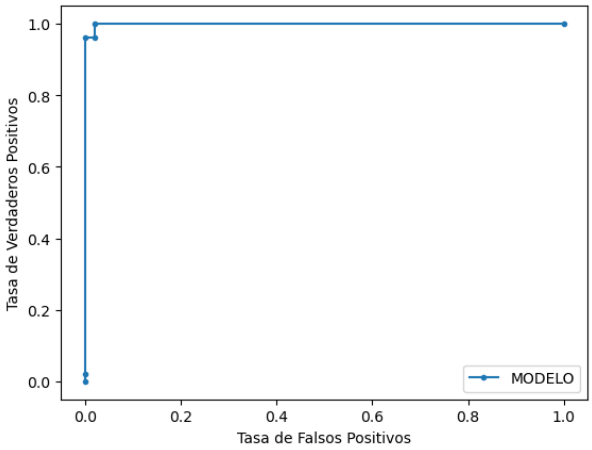

Figura 16: Curva ROC de validación de fonemas (Autoría propia)

En base a la Figura 15 el modelo entrenado con espectrograma de fonemas tiene un 95% de precisión. Además, en la curva ROC de la Figura 16 notamos que la curva está bastante por encima de la diagonal principal, mostrando que el modelo es un buen clasificador.

3.3.5.2 Matriz de confusión utilizando la lectura de texto

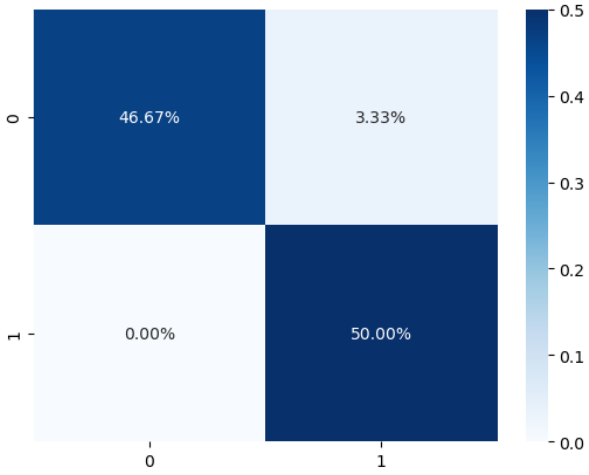

Figura 17: Matriz de confusión de validación de texto (Autoría propia)

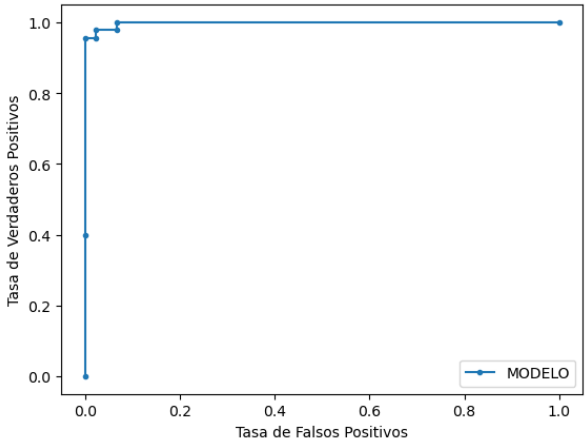

Figura 18: Curva ROC de validación de texto (Autoría propia)

Según la Figura 17, el modelo entrenado con espectrogramas de audios de texto tiene una precisión de 96.67%, siendo mejor que el modelo de fonemas. Además, en la Figura 18 notamos que la curva está bastante por encima de la diagonal principal, aunque no tanto como el clasificador de fonemas, aun así, el clasificador sigue siendo bueno.

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

En el desarrollo de este trabajo, se utilizaron tres fuentes de archivos de audio, los cuales fueron obtenidos del Centro Integral en Neurociencias – CINAC, de HM Hospitales de la Universidad San Pablo y de pacientes y sujetos sanos locales. Estos datos fueron usados para entrenar y validar los modelos de inteligencia artificial utilizados en los diversos experimentos realizados con modelos Deep Learning y Machine Learning.

Previo al entrenamiento de los modelos de inteligencia artificial, se pre-procesaron los archivos de audio, recortándolos y realizando un aumento de datos usando la técnica de “slide window”, con lo que logramos incrementar la cantidad de audios a utilizar. Una vez realizado el pre-procesamiento, se realizó la extracción de características utilizadas para los modelos de Machine Learning, y la generación de imágenes de espectrogramas para los modelos de Deep Learning.

Los modelos de Machine Learning utilizados para el análisis de características fueron Support Vector Machine (SVM), XGBoost, Multi Layer Perceptron (MLP), Random Forest y Regresión Logística. Mientras que los modelos de Deep Learning utilizados para el análisis de espectrogramas fueron CNN, ResNet 50 y ResNet 50 v2. Todos estos modelos fueron evaluados para determinar el mejor modelo para el producto final.

Se desarrolló un prototipo web donde se integró el modelo final, el cual permite a un doctor realizar un diagnóstico rápido de la EP, brindando un indicio de si el sujeto pueda o no padecer de la enfermedad.

4.1 Conclusiones

- Se desarrollaron dos modelos usando técnicas de inteligencia artificial para el análisis de espectrogramas. Se usó la arquitectura convolucional ResNet 50 v2 para generar un modelo para el análisis de fonemas y otro para el análisis de textos. Ambos modelos analizan audios de dos segundos con un rango de frecuencias de 0 a 8000 Hz, obteniendo una precisión de 95% en el análisis de fonemas y 97% en el análisis de textos.

- Se grabó a pacientes con Parkinson de la clínica Kennedy en Guayaquil, y a sujetos sanos cercanos a los autores de este documento. Esta información se añadió a las grabaciones enviadas por la clínica neurologíca en España.
- Se evaluaron características de la voz obtenidas con OpenSMILE y el uso de espectrogramas de la voz. Finalmente, solo se utilizó los espectrogramas de la voz debido a que presentaron mejores resultados a comparación de las características de la voz.
- Finalmente, se desarrolló una aplicación web que usa el mejor modelo entrenado para evaluar audios de personas y determinar la probabilidad de que padezca de la enfermedad o no. Esta aplicación fue desarrollada para dos tipos de usuario: el doctor quien realizará el análisis y el administrador del sistema. Las funcionalidades definidas para el doctor consisten en la subida de audios e información del sujeto para su diagnóstico usando el modelo entrenado. También el sistema admite la grabación de audios y la visualización del historial de diagnósticos. Mientras que el administrador cuenta con los permisos para crear, actualizar y eliminar registros dentro de la base de datos definida para el prototipo.

4.2 Recomendaciones

Se recomienda seguir un protocolo de grabación de los audios para la página web, debido a que, si no se graba en buenas condiciones, el modelo podría diagnosticar que el audio ingresado pertenece a un paciente, cuando en realidad es un sujeto sano con malas condiciones de grabación.

Además, se recomienda usar este modelo bajo la supervisión de un profesional, ya que la enfermedad de Parkinson debe ser evaluada usando información de las diferentes sintomatologías, con lo que este modelo abre una nueva opción para el análisis de voz.

Un posible uso comercial de estos modelos se puede dar en aplicaciones que evalúan constantemente el estado de salud de una persona. O incluso lograr añadir esta funcionalidad en asistentes personales como Amazon Alexa o Google Assistant, donde se hace uso de la voz para poder enviar instrucciones, donde de manera pasiva, si el usuario lo desea, se evalúe la potencia de la voz y en caso de

que el modelo muestre una alta probabilidad de padecer la enfermedad, el asistente recomendaría visitar a un médico.

BIBLIOGRAFÍA

[1] World Health Organization, «Parkinson disease», *World Health Organization*, 13 de junio de 2022. https://www.who.int/news-room/fact-sheets/detail/parkinson-disease (accedido 17 de octubre de 2022).

[2] C. Marras *et al.*, «Prevalence of Parkinson's disease across North America», *NPJ Park. Dis.*, vol. 4, n.º 1, pp. 1-7, 2018.

[3] P. Pastor y E. Tolosa, «La enfermedad de Parkinson: diagnóstico y avances en el conocimiento de la etiología y en el tratamiento», *Med. Integral*, vol. 37, n.º 3, pp. 104-117, 2001.

[4] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins, y R. H. Ghomi, «Parkinson's disease diagnosis using machine learning and voice», en *2018 IEEE signal processing in medicine and biology symposium (SPMB)*, 2018, pp. 1-7.

[5] A. S. Gullapalli y V. K. Mittal, «Early detection of Parkinson's disease through speech features and machine learning: a review», *ICT Intell. Appl.*, pp. 203-212, 2022.

[6] B. Jin, Y. Qu, L. Zhang, Z. Gao, y others, «Diagnosing Parkinson disease through facial expression recognition: video analysis», *J. Med. Internet Res.*, vol. 22, n.º 7, p. e18697, 2020.

[7] M. Fabbri, L. A. Kauppila, J. J. Ferreira, y O. Rascol, «Challenges and perspectives in the management of late-stage Parkinson's disease», *J. Park. Dis.*, vol. 10, n.º s1, pp. S75-S83, 2020.

[8] R. L. Nussbaum y C. E. Ellis, «Alzheimer's disease and Parkinson's disease», *N. Engl. J. Med.*, vol. 348, n.º 14, pp. 1356-1364, 2003.

[9] R. Martínez-Fernández, Á. Sánchez-Ferro, J. Á. Obeso, y others, «Actualización en la enfermedad de Parkinson», *Rev. Médica Clínica Las Condes*, vol. 27, n.º 3, pp. 363-379, 2016.

[10] P. Pollak *et al.*, «Treatment results: Parkinson's disease», *Mov. Disord. Off. J. Mov. Disord. Soc.*, vol. 17, n.º S3, pp. S75-S83, 2002.

[11] B. S. Connolly y A. E. Lang, «Pharmacological treatment of Parkinson disease: a review», *Jama*, vol. 311, n.º 16, pp. 1670-1683, 2014.

[12] L. Brabenec, J. Mekyska, Z. Galaz, y I. Rektorova, «Speech disorders in Parkinson's disease: early diagnostics and effects of medication and brain stimulation», *J. Neural Transm.*, vol. 124, n.º 3, pp. 303-334, 2017.

[13] M. Picó Berenguer y H. A. Yévenes Briones, «SPEECH DISORDERS IN PARKINSON'S DISEASE. REVIEW», *Rev. Científica Cienc. Médica*, vol. 22, n.º 1, pp. 36-42, 2019.

[14] M. Little, P. McSharry, E. Hunter, J. Spielman, y L. Ramig, «Suitability of dysphonia measurements for telemonitoring of Parkinson's disease», *Nat. Preced.*, pp. 1-1, 2008.

[15] F. Eyben, M. Wöllmer, y B. Schuller, «Opensmile: the munich versatile and fast open-source audio feature extractor», en *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459-1462.

[16] S. Lahmiri, D. A. Dawson, y A. Shmuel, «Performance of machine learning methods in diagnosing Parkinson's disease based on dysphonia measures», *Biomed. Eng. Lett.*, vol. 8, n.º 1, pp. 29-39, 2018.

[17] W. Caesarendra, F. T. Putri, M. Ariyanto, y J. D. Setiawan, «Pattern recognition methods for multi stage classification of Parkinson's disease utilizing voice features», en *2015 IEEE International Conference on Advanced Intelligent Mechatronics (AIM)*, 2015, pp. 802-807.

[18] S. Lahmiri y A. Shmuel, «Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine», *Biomed. Signal Process. Control*, vol. 49, pp. 427-433, 2019.
[19] B. E. Boser, I. M. Guyon, y V. N. Vapnik, «A training algorithm for optimal margin classifiers», en *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144-152.
[20] N. Cristianini, J. Shawe-Taylor, y others, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
[21] I. T. Jolliffe y J. Cadima, «Principal component analysis: a review and recent developments», *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 374, n.º 2065, p. 20150202, 2016.
[22] G. A. Martinez Mascorro y G. Aguilar Torres, «Voice recognition based on MFCC, SBC and Spectrograms», *INGENIUS-Rev. Cienc. Tecnol.*, n.º 10, pp. 12-20, 2013.
[23] I. Hugo Olmedo, «Análisis espectrográfico del habla en pacientes afectados por párkinson para contribuir al diagnóstico», Universidad de Cádiz, 2018. [En línea]. Disponible en: https://rodin.uca.es/bitstream/handle/10498/20675/Trabajo%20de%20Fin%20de%20Grado.pdf?sequence=2&isAllowed=y#:~:text=El%20habla%20de%20los%20pacientes,S%C3%A1nchez%2C%202010%3A%20542).].
[24] X. Zou, Y. Hu, Z. Tian, y K. Shen, «Logistic regression model optimization and case analysis», en *2019 IEEE 7th international conference on computer science and network technology (ICCSNT)*, 2019, pp. 135-139.
[25] T. Chen y C. Guestrin, «Xgboost: A scalable tree boosting system», en *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785-794.
[26] H. Ramchoun, Y. Ghanou, M. Ettaouil, y M. A. Janati Idrissi, «Multilayer perceptron: Architecture optimization and training», 2016.
[27] Y. Kong y T. Yu, «A deep neural network model using random forest to extract feature representation for gene expression data classification», *Sci. Rep.*, vol. 8, n.º 1, pp. 1-9, 2018.
[28] J. R. Brzezinski y G. J. Knafl, «Logistic regression modeling for context-based classification», en *Proceedings. Tenth International Workshop on Database and Expert Systems Applications. DEXA 99*, 1999, pp. 755-759.
[29] K. O'Shea y R. Nash, «An introduction to convolutional neural networks», *ArXiv Prepr. ArXiv151108458*, 2015.
[30] K. He, X. Zhang, S. Ren, y J. Sun, «Deep residual learning for image recognition», en *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
[31] O. Caelen, «A Bayesian interpretation of the confusion matrix», *Ann. Math. Artif. Intell.*, vol. 81, n.º 3-4, pp. 429-450, 2017.
[32] M. Gönen y others, «Receiver operating characteristic (ROC) curves», *SAS Users Group Int. SUGI*, vol. 31, pp. 210-231, 2006.

Anexos

ANEXO 1

Aplicación de escritorio para grabación de audios.

Esta aplicación fue desarrollada en Python y permite realizar grabaciones de audios con el dispositivo de entrada que se especifique. Esta aplicación fue la usada para realizar las grabaciones a los pacientes y sujetos de control.

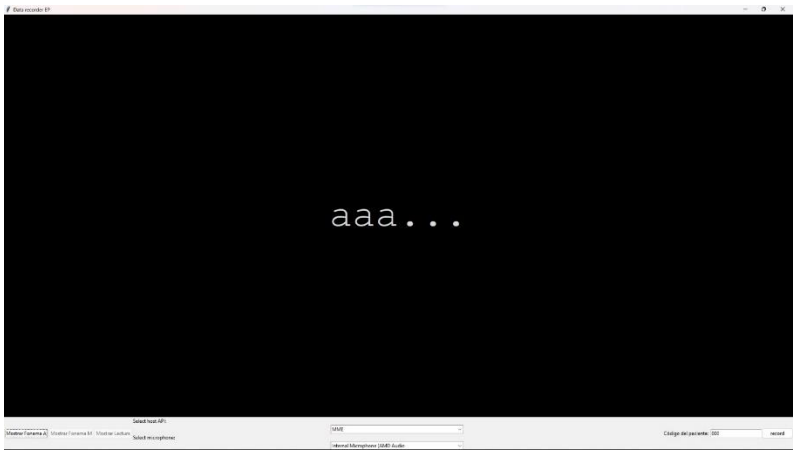

Figura A 1: Grabación del fonema A.

Figura A 2: Grabación del fonema M.

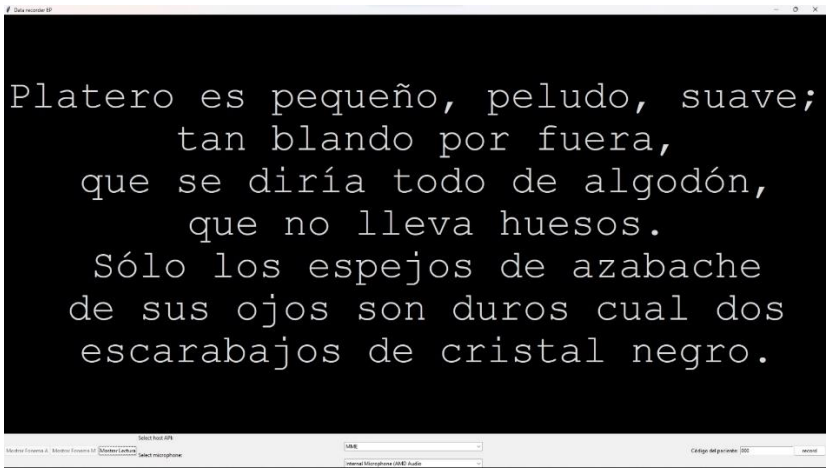

Figura A 3: Grabación de la lectura del texto.