Universidad Nacional
de La Matanza

ESCUELA DE POSGRADO

TESIS DE MAESTRÍA EN INFORMÁTICA

Detección de la Enfermedad de Parkinson basada en Máquinas de Aprendizaje Extremo

Autora: Lic. Renata Silvia Guatelli

Directora: Dra. Verónica Inés Aubin

Buenos Aires. Agosto 2023.

Página intencionalmente en blanco

Agradecimientos

A mis padres.

A mi directora, la Dra. Verónica Inés Aubin quien supo guiarme en la confección de esta tesis.

Al Dr. Marco Mora Cofré, por el acceso al Laboratorio de Investigaciones Tecnológicas en Reconocimiento de Patrones, de la Universidad Católica del Maule, Chile, donde se desarrollaron los experimentos de esta tesis, además de sus invaluables comentarios y consejos.

Al Ing. Jorge Doorn y a su equipo de investigación, en especial a la Dra. Gladys Kaplan, quienes me guiaron en mis primeros pasos en la investigación.

A mi comadre, quien acompañó todo el proceso de desarrollo de esta tesis, sufrió mi desesperanza e inseguridades y supo darme confianza para poder culminar este trabajo.

A mis colegas y amigas quienes me acompañaron durante todo el proceso, brindando su apoyo y consejos valiosos.

A Denise y Ana quienes leyeron y aportaron comentarios para mejorar este trabajo.

A la Universidad de la Matanza por darme la oportunidad de ser docente de la institución y formarme como investigadora.

Página intencionalmente en blanco

Resumen

La enfermedad de Parkinson es un trastorno del movimiento que se caracteriza por la degeneración de las células nerviosas en una región del cerebro llamada sustancia negra mesencefálica, afectándose las vías generadoras de dopamina, un neurotransmisor esencial para el control del movimiento. Sus causas son variadas, dentro de ellas se encuentra la exposición a pesticidas, factores genéticos y, uno de los más influyentes, la edad. Si bien aún no se conoce una cura para la enfermedad, existen tratamientos que pueden mejorar significativamente la calidad de vida de los pacientes. Dado la disminución de dopamina, los síntomas más comunes son la aparición de temblores y rigidez muscular. Debido a la rigidez de los músculos se producen alteraciones de la voz las cuales tienen gran potencial para el diagnostico no invasivo y precoz de la enfermedad. El bajo costo de este diagnostico en comparación con los estudios clínicos lo haría accesible a un mayor número de personas. Trabajos recientes que analizan grabaciones de voz mediante Redes Neuronales Convolucionales presentan elevados niveles de acierto en el diagnóstico de la Enfermedad de Parkinson. En esta tesis se presentan distintos modelos de aprendizaje profundo, para la clasificación de enfermos de Parkinson y no enfermos utilizando espectrogramas de las señales de voz. Para construir clasificadores con Redes Neuronales Convolucionales, se requiere un número elevado de muestras y largo tiempo de entrenamiento. Para superar estas desventajas en esta tesis se presentan dos estrategias de aumentación de datos, a) la generación de espectrogramas con distintas paletas de colores; b) la fragmentación del audio original en segmentos de 1 segundo con el 50 % de solapamiento y el uso de Máquinas de Aprendizaje Extremo aplicados a espectrogramas de voz. Se presentan 4 experimentos aplicados a espectrogramas con diversas arquitecturas de Redes Neuronales Convolucionales, a saber, AlexNet, VGG-16, SqueezeNet, Inception V3 y ResNet-50. En los experimentos se compara objetivamente la tasa de acierto, el tiempo de entrenamiento y test, la sensibilidad y la especificidad de todas las arquitecturas neuronales involucradas en el trabajo. Se muestra que las Máquinas de Aprendizaje Extremo tienen un nivel elevado de acierto en el diagnóstico de la enfermedad de Parkinson pero con tiempos de entrenamiento reducidos.

Página intencionalmente en blanco

Abstract

Parkinson's disease is a movement disorder characterized by the degeneration of nerve cells in a region of the brain called the substantia nigra mesencephalic, affecting the pathways that generate dopamine, an essential neurotransmitter for the control of movement. motion. Its causes are varied, among them is exposure to pesticides, genetic factors and, one of the most influential, age. Although there is still no known cure for the disease, there are treatments that can significantly improve the quality of life of patients. Given the decrease in dopamine, the most common symptoms are the appearance of tremors and muscle stiffness. Due to the rigidity of the muscles, voice alterations occur, which have great potential for non-invasive and early diagnosis of the disease. The low cost of this diagnosis compared to clinical studies would make it accessible to a greater number of people. Recent works that analyze voice recordings using Convolutional Neural Networks present high levels of accuracy in the diagnosis of Parkinson's Disease. In this thesis, different deep learning models are presented for the classification of Parkinson's patients and non-patients using spectrograms of voice signals. To build classifiers with Convolutional Neural Networks, a large number of samples and a long training time are required. To overcome these disadvantages, in this thesis two data augmentation strategies are presented: a) the generation of spectrograms with different color palettes; b) the fragmentation of the original audio into segments of 1 second with 50 % overlap and the use of Extreme Machine Learning applied to voice spectrograms. Four experiments applied to spectrograms with various Convolutional Neural Network architectures are presented, namely AlexNet, VGG-16, SqueezeNet, Inception V3 and ResNet-50. In the experiments, the success rate, the training and test time, the sensitivity and the specificity of all the neural architectures involved in the work are objectively compared. It is shown that Extreme Learning Machines have a high level of accuracy in the diagnosis of Parkinson's disease but with reduced training times.

Contenido

Página intencionalmente en blanco

Lista de figuras

Lista de tablas

Página intencionalmente en blanco

Capítulo 1

Introducción

1.1. Introducción

La enfermedad de Parkinson (EP) es un desorden neurodegenerativo del sistema nervioso, de causa desconocida, de curso crónico, progresivo e irreversible. Es la segunda enfermedad neurodegenerativa más frecuente después del Alzheimer. Los primeros síntomas no se hacen visibles hasta pasados varios años de padecer la enfermedad. Las manifestaciones motoras como el temblor son las más conocidas. Sin embargo, existen otros síntomas que en muchas ocasiones son invalidantes [9]. Entre estos se destacan problemas cognitivos, rigidez muscular, inestabilidad postural y lentitud del movimiento [10, 11]. El diagnóstico clínico de la EP se basa en el seguimiento de la historia clínica del paciente junto con análisis físicos y neurológicos para determinar la presencia de algunos síntomas o la ausencia de otros. El único diagnóstico definitivo es el histopatológico. Es decir, el que se realiza durante la autopsia [12]. En los últimos años diferentes investigaciones proponen sistemas de diagnóstico para la detección temprana de la enfermedad al considerar diferentes biomarcadores (voz, escritura, marcha, etc.). Estas contribuyen con la identificación de la enfermedad en estadios tempranos, beneficiando al paciente con la posibilidad de recibir un tratamiento antes que la pérdida de neuronas sea mayor [13].

Debido a la rigidez de la musculatura, los enfermos de Parkinson comienzan

a presentar distorsiones en la voz. En las voces patológicas y particularmente en los enfermos de Parkinson se observan variaciones temporales mayores de la intensidad o amplitud de la señal. Medir las distorsiones o cambios en la voz proporciona información útil para diferenciar entre voces patológicas y voces sanas [14, 15], así como para identificar mediante esquemas de clasificación el estadío de la enfermedad [16]. Los sistemas de diagnóstico basados en el análisis de la voz presentan interesantes ventajas pues permiten el diagnóstico no invasivo y precoz de la enfermedad [15]. El estudio de la voz aplicado a esta problemática se aborda principalmente desde dos enfoques. Un enfoque busca estudiar distintos algoritmos para extraer características de la voz, y luego utiliza dichas características como entrada de clasificadores [17]. El otro enfoque emplea el aprendizaje profundo sin buscar descriptores, y trabaja directamente en el dominio original de la información (señales o espectrogramas de la voz). Un espectrograma es una representación visual de la señal de sonido que considera el tiempo, la frecuencia y la amplitud [18]. La representación del habla a través de espectrogramas ha demostrado ser estable y robusta, incluso en presencia de elevados niveles de ruido [19, 20].

En los últimos años, las técnicas de aprendizaje profundo han demostrado ser efectivas en la resolución de problemas de clasificación en diversas áreas. Diversos autores han utilizado las redes neuronales convolucionales (CNN) para abordar el diagnóstico de la EP [21, 22, 23, 24]. No obstante, las CNN son una solución efectiva si se dispone de grandes conjuntos de datos (en el entrenamiento se requiere estimar millones de parámetros). Por otro lado, es necesario tener servidores de cómputo equipados con procesadores de múltiples núcleos, y disponibilidad de aceleradores como las unidades de procesamiento gráfico (GPU), para disminuir el excesivo tiempo de entrenamiento requerido por las CNN [25]. Además, el algoritmo que principalmente se usa para entrenar una CNN, llamado Backpropagation (BP), es lento [26]. Por último, una cuestión importante a considerar es que en este problema se cuenta con un número reducido de muestras para analizar, pues las bases de datos disponibles tienen muy

pocos pacientes (esto se explica porque el protocolo de adquisición de muestras es complejo, intervienen muchos especialistas y de alto costo). Para subsanar el problema de contar con pocas muestras, una de las técnicas es aplicar la transferencia de aprendizaje. Sin embargo, el proceso de entrenamiento aún es costoso en términos de tiempo.

Las Máquinas de Aprendizaje Extremo (ELM) corresponden a un paradigma de entrenamiento para Redes Neuronales supervisadas [27], semisupervisadas [28] y no supervisadas [29]. El entrenamiento de las ELM consiste básicamente en asignar aleatoriamente el valor de los parámetros de las capas ocultas, y reducir el entrenamiento a la estimación de los parámetros de la capa de salida mediante la resolución de un sistema de ecuaciones lineales sobredeterminado utilizando la Pseudoinversa de Moore-Penrose [30, 31]. Recientemente, la literatura presenta la aplicación de las ELM a problemas de clasificación del área de la salud como: enfermedad de Alzheimer [32], Parkinson utilizando parámetros acústicos [33], epilepsia [34], cáncer de mama [35], lesiones de piel [36], entre otras.

La línea de investigación de esta tesis ha sido discernir entre pacientes con EP y personas sin EP a partir del uso de técnicas de aprendizaje profundo aplicadas a los espectrogramas de señales de voz obtenidas a partir de grabaciones realizadas en un ambiente controlado, comparando las ventajas del uso de ELM sobre las CNN.

1.2. Dominio del Problema

El promedio de esperanza de vida viene aumentando en forma sostenida a nivel mundial, gracias a los avances médicos y científicos que permitieron lograr mejores condiciones de infraestructura, prácticas de higiene y tratamientos médicos. Una de las características negativas asociadas a este incremento en la esperanza de vida es el mayor número de enfermedades neurodegenerativas, incluido el Parkinson. La Asociación Europea de la Enfermedad de Parkinson

estima que, para el año 2030, entre 8.7 y 9.3 millones de personas sufrirán esta enfermedad [37].

Esto representa un serio problema de salud pública. Aunque el curso de la EP es irreversible, existen terapias que permiten que los pacientes vean reducidos los síntomas de la enfermedad, lo que les permite lograr una mejor calidad de vida cotidiana y autónoma cuanto antes se detecte la enfermedad.

La detección temprana de la EP es una tarea difícil dado que los síntomas de la enfermedad no son visibles hasta al menos cuatro años después de su inicio. Algunas investigaciones previas han demostrado que los pacientes que se encuentran en una etapa temprana de la misma tienen algún trastorno en la voz. Diversos estudios ya han propuesto la utilización del procesamiento de la voz para obtener parámetros acústicos como método objetivo y no invasivo para valorar síntomas iniciales en EP. La utilización de técnicas de Deep Learning aplicadas a espectrogramas de señales de audio es un aspecto considerado en algunos trabajos publicados sobre la detección de la EP, pero aún tiene un potencial importante por explotar.

1.3. Hipótesis

Es posible mejorar el tiempo de entrenamiento y clasificación de las CNN para la detección de la enfermedad de Parkinson, utilizando Máquinas de Aprendizaje Extremo, manteniendo los niveles de exactitud (accuracy).

1.4. Objetivos

1.4.1. Objetivo general

Considerando espectrogramas obtenidos a partir de señales de voz, el objetivo general de este trabajo consiste en reducir el tiempo de entrenamiento y clasificación de las redes neuronales convolucionales para la detección de la enfermedad de Parkinson, mediante el uso de Máquinas de Aprendizaje Extremo.

1.4.2. Objetivos específicos

- Desarrollar un repositorio de espectrogramas de las señales de voz.
- Desarrollar técnicas de aumentación de datos para el adecuado entrenamiento de las redes neuronales.
- Desarrollar un modelo de detección de la enfermedad de Parkinson con Máquinas de Aprendizaje Extremo.
- Comparar los modelos de redes neuronales convolucionales y Máquinas de Aprendizaje Extremo.

1.5. Organización de la Tesis

En el capítulo 1 se presenta la motivación, hipótesis, objetivos principales de esta tesis y una breve descripción del dominio del problema. En el capítulo 2 se presenta el marco teórico, donde se exponen aspectos relacionados con el Parkinson y la voz y los fundamentos de los modelos de Deep Learning. Se expone la revisión del estado del arte realizada para el presente estudio. En el capítulo 3 se detallan los métodos aplicados, los materiales utilizados y los experimentos realizados en búsqueda de cumplir con los objetivos planteados. En el capítulo 4 se muestran los resultados de los experimentos realizados y se analizan tomando como referencia los objetivos planteados. Finalmente, en el capítulo 5 se exponen las conclusiones del trabajo realizado y se proponen futuras líneas de investigación que pueden plantearse a partir de esta tesis.

Página intencionalmente en blanco

Capítulo 2

El Problema: El Parkinson y la Voz

2.1. Enfermedad de Parkinson

2.1.1. Antecedentes

El nombre de la enfermedad se debe a James Parkinson, el médico que en 1817 hace una descripción completa y minuciosa de la "Shaking Palsy" o "Parálisis temblorosa" [38]. El mérito de Parkinson radicó en integrar en un único trastorno manifestaciones que hasta ese momento eran consideradas enfermedades diferentes.

2.1.2. Definición

La EP es un trastorno degenerativo crónico del sistema nervioso central, asociada con una disminución progresiva de las habilidades motoras y la integración sensoriomotora [39], de progresión lenta y, por el momento, irreversible. También se la denomina "parkinsonismo primario" o "Enfermedad de Parkinson idiopática". El término idiopático en medicina hace referencia a que las causas que provocan la enfermedad son desconocidas. Es un trastorno neurodegenerativo que consiste en la degeneración de los cuerpos de Lewy y mayormente de la zona denominada Sustancia Negra, que se localiza en la parte del cerebro conocida como "mesencéfalo". Esta región se considera parte del tronco cerebral [40]. Hay una sustancia negra en cada hemisferio del encéfalo y tal

sustancia es la encargada de coordinar los movimientos, la actividad y el tono muscular. Las neuronas de esta zona son las encargadas de producir la dopamina, que es el principal neurotransmisor para la función motora del organismo [41, 42]. Sin los niveles adecuados de dopamina aparecen síntomas como el temblor, la rigidez, la lentitud de movimiento y la inestabilidad postural.

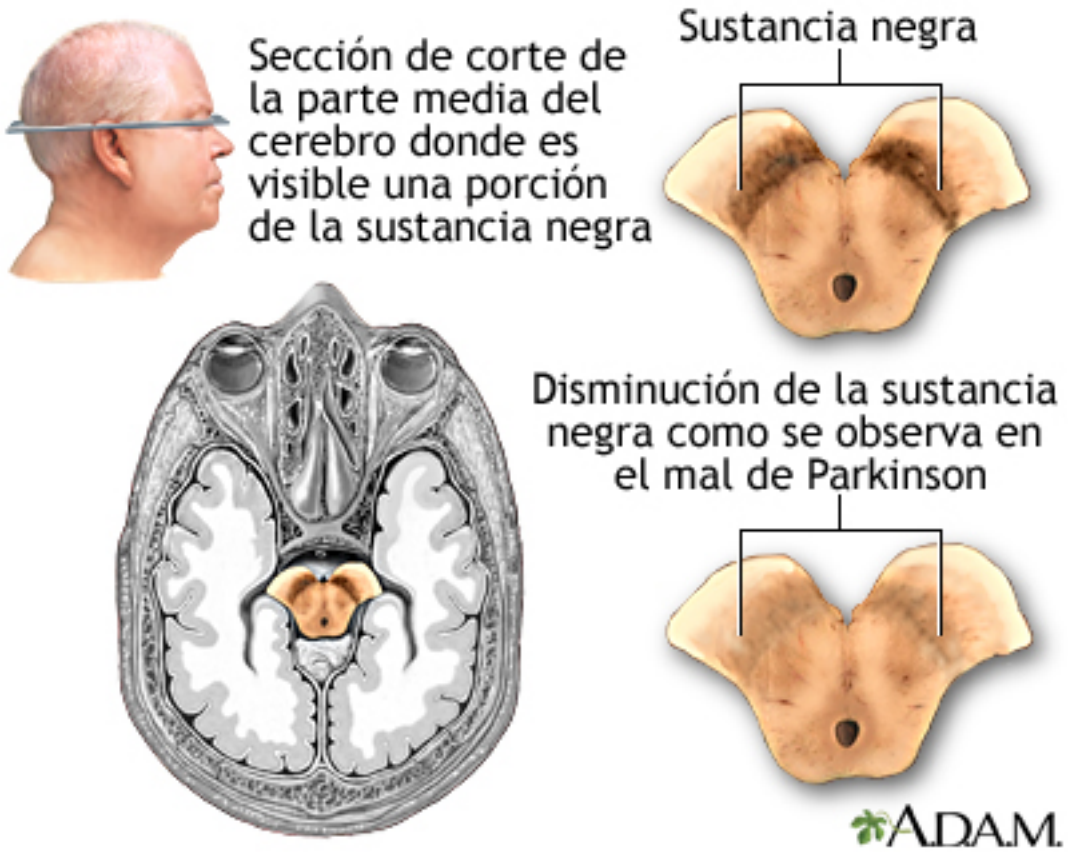


Figura 1: Sustancia negra y la enfermedad de Parkinson. Fuente [1]

Es una enfermedad crónica, lo cual implica que persiste hasta el final de la vida, ya que, hasta este momento, no se ha encontrado una cura. Es progresiva, por lo cual sus síntomas empeoran a medida que transcurre el tiempo de padecer la enfermedad.

2.1.3. Síntomas

Las manifestaciones clínicas de los pacientes con EP se pueden clasificar en: motores y no motores [43, 44]. La EP se cataloga como perteneciente al grupo de enfermedades de “trastornos del movimiento”. Los síntomas principales, son

los síntomas motores: temblor especialmente en manos, brazos, piernas mandíbula y/o cabeza; rigidez de las extremidades y tronco; bradicinesia (lentitud en los movimientos) e inestabilidad postural (deterioro del equilibrio). Entre los síntomas no motores se pueden mencionar: depresión, ansiedad, alucinaciones, astenia (cansancio), dolores musculares generalizados o localizados resistentes a tratamientos analgésicos, pérdida del olfato total o parcial, dificultades en la visión (dificultad para detectar colores, ojos secos, visión doble), babeo (la saliva se acumula en la boca y como hay pérdida de movimiento se dificulta su deglución), disfagia (debido a la falta de coordinación para tragar), estreñimiento, disfunción urinaria, disfunción eréctil, trastornos del sueño, disfunción sexual, sudoración excesiva, cambios de peso [45, 46]. La literatura indica que estos síntomas aparecen varios años antes de que se puedan detectar los síntomas motores. A medida que avanza la enfermedad los síntomas se agudizan pudiendo generar dificultades para caminar, hablar o realizar tareas comunes, estos síntomas degradan enormemente la calidad de vida del paciente y en muchas ocasiones se vuelven invalidantes [11]. No todos los pacientes que tienen estos síntomas padecen la enfermedad. El diagnóstico de temblor de reposo, rigidez y bradicinesia caracterizan a un amplio grupo de enfermedades, que se conoce como síndrome parkinsoniano o parkinsonismo [47]. Solo el 20 % de los pacientes con estos síntomas padecen alguna otra enfermedad neurodegenerativa [48].

2.1.4. Prevalencia e Incidencia

La edad de inicio promedio es de 60 años. A partir de esta edad la prevalencia de la enfermedad (frecuencia de casos en una población y en un momento dado) aumenta exponencialmente. Si la enfermedad comienza antes de los 50 años se la denomina “EP de inicio temprano”. En general, se asocia a variantes genéticas. En este caso, los síntomas motores suelen aparecer entre los 21 a 45 años. Esta variante representa alrededor del 5 % de los pacientes con EP [49]. En términos de fallecimientos y discapacidades, la carga global de la enfermedad,

se ha más que duplicado en las últimas dos décadas [50]. Tanto la incidencia (probabilidad de que una persona de una cierta población resulte afectada por dicha enfermedad) como la prevalencia de la EP es de 1,5 a 2 veces mayor en hombres que en mujeres [51, 52].

2.1.5. Evolución de la Enfermedad

Según la Sociedad Internacional de Parkinson y Trastornos del Movimiento se pueden considerar tres etapas en la evolución natural de la EP: 1) Una fase preclínica, donde los procesos neurodegenerativos ya han comenzado pero aún no existen síntomas o signos clínicos visibles; 2) Una fase prodrómica, en la que algunos de los síntomas y signos son detectados pero aún no son suficientes para el diagnóstico de la enfermedad; y 3) Una fase clínica, en la que es posible diagnosticar la EP basado en la presencia de signos motores [53]. La fase prodrómica suele comenzar varios años antes de que los síntomas motores puedan ser detectados. Para medir la evolución de la enfermedad, tanto en ámbitos clínicos como de investigación, se utiliza la "Escala unificada de la enfermedad de Parkinson modificada por la MDS" (MDS-UPDRS) [54], la cual realiza un abordaje completo e integral de los aspectos clínicos relevantes de la EP.

2.1.6. Diagnóstico de la Enfermedad

El diagnóstico clínico de la EP se basa en el seguimiento de la historia clínica del paciente junto con análisis físicos [55, 56] y neurológicos [57, 58] para determinar la presencia o ausencia de determinados síntomas. En 2015, el "Movement Disorder Society Clinical Diagnostic Criteria for Parkinson's disease" [59] establece los criterios clínicos para declarar la EP. Estos criterios han demostrado ser confiables ya que son ampliamente utilizados a nivel mundial, y son aceptados como estándar en ensayos clínicos y proyectos de investigación. No obstante lo anterior, el único diagnóstico definitivo es el histopatológico. Es decir, el que se realiza durante la autopsia [60, 12]. La enfermedad generalmente se diagnostica luego de la aparición de síntomas motores. Estos síntomas apare-

cen cuando se han perdido aproximadamente el 80 % de las células dopaminérgicas [61]. La destrucción de neuronas no es simétrica, lo que hace que, inicialmente, los síntomas se hagan visibles en forma unilateral y progresivamente se vuelvan bilaterales. Al diagnosticar la enfermedad en etapas tardías, las terapias tienen pocas probabilidades de detener el progreso de la enfermedad. Teniendo en cuenta esto, el diagnóstico temprano es esencial para el aumento de la neuroprotección y mejoría en el pronóstico [62]. Sin embargo, dado que existen patologías con síntomas similares, su actual tasa diagnóstica es insatisfactoria y su detección temprana requiere de una mayor complejidad [63].

2.1.7. Detección y Seguimiento de la Enfermedad de Parkinson

En las últimas décadas, se han desarrollado diversos biomarcadores para el estudio de los pacientes con EP. Algunos de estos estudios monitorean los patrones de voz, dada su relevancia para esta investigación estos se desarrollaran en 2.2, mientras que otros que analizan la escritura a mano [64, 65, 66, 67, 68], como una herramienta eficaz para el diagnóstico temprano de la EP. Kotsavasiloglou en [69], explora el uso de un dispositivo de bolígrafo y tableta para estudiar las diferencias entre el movimiento de las manos y los músculos de sujetos sanos y pacientes con EP al escribir. Pereira en [70], realiza la identificación automática de EP mediante información dinámica de la escritura manuscrita. Drotár en [66], presenta la base de datos PaHaW formada por imágenes que corresponden a dibujos de una espiral arquimediana, la escritura repetitiva de sílabas y palabras ortográficamente simples y una oración. Otros estudios consideran biomarcadores, tales como la marcha [71] y el movimiento al pulsar los botones del teclado de las computadoras [72]. Otro grupo de biomarcadores considerados para el diagnóstico de la EP son las neuroimágenes, como la resonancia magnética cerebral [73] y las imágenes moleculares, en particular, la tomografía por emisión de positrones (PET) [74] y la tomografía por emisión de fotón único (SPECT) [75].

2.2. La Voz Humana

La voz es una característica fundamental que nos ha permitido evolucionar como especie y es la base de la comunicación oral. Thomas Henry Huxley [76] ya la resaltó en 1871 como el carácter distintivo del hombre. En la actualidad, el estudio del habla y de la voz se aborda desde distintos campos, como la síntesis de voz, el reconocimiento de voz, el reconocimiento del hablante y el análisis de las características acústicas de la voz. En el siguiente texto se describirán brevemente estas características.

2.2.1. Fisiología de la Voz

La voz humana es producida en la laringe. Sin embargo, prácticamente todos los sistemas del cuerpo la afectan. El aire procedente de los pulmones es forzado durante la espiración a través de la glotis, haciendo vibrar los dos pares de cuerdas vocales. Para generar la voz humana, entran en funcionamiento los órganos de respiración (pulmones, bronquios y tráquea); los órganos de fonación y resonancia (laringe, cuerdas vocales y resonadores nasal, bucal y faríngeo); además de los órganos de articulación (paladar, lengua, dientes, labios y glotis) [77, 78].

2.2.2. Análisis Acústico

Las medidas físicas de la voz humana se basan en el empleo de diversos parámetros acústicos que reflejan las tres dimensiones perceptibles del sonido: amplitud, tono y estructura temporal. La amplitud (cuyo principal parámetro es la intensidad) es una medida de la presión sonora al transmitirse la voz en el medio aéreo, expresada en decibelios (dB), dependiente de la amplitud de la vibración de las cuerdas vocales y de la presión subglótica [79], el tono se expresa mediante la F0 de la señal vocal, medida en hertzios (Hz); por último, los parámetros derivados del tiempo tienen relación con la tasa y rapidez de la vocalización. La representación de las variables físicas es altamente compleja a

causa de la variabilidad de la energía espectral de la voz. La frecuencia fundamental (F0) es la principal unidad de análisis acústico.

2.2.3. Algunas Medidas Físicas de la Voz Humana

2.2.3.1. La Frecuencia Fundamental (F0)

Es el número de veces que vibran los pliegues vocales por segundo. Es la principal unidad de análisis acústico. Sus valores varían dependiendo del sexo y la edad. Los niños, sin importar el sexo, tienen una frecuencia media de 240 Hz hasta la pubertad, en la que los varones tienen un descenso hasta los 110 Hz (la voz se torna más grave), mientras que las mujeres se mantienen en 210 Hz. Hacia la tercera edad, aumenta la frecuencia de los hombres a 140 Hz y disminuye en las mujeres a 190 Hz en promedio [80]. Estos valores se modifican en la voz senil, reduciéndose la F0 en las mujeres a 175 Hz aproximadamente, mientras que se incrementa en el hombre a 130 Hz a los 70 años y a 160 Hz a los 90 años .

2.2.3.2. Relación Armónico-Ruido (HNR)

La relación armónico ruido HNR (Harmonics-to-Noise-Ratio) es una medida de la proporción de sonido armónico a ruido en la voz, medida en decibelios [81]. El sonido producido por las vibraciones de los pliegues vocales está compuesto por ondas sonoras periódicas y aperiódicas. Las ondas aperiódicas son ruido aleatorio introducido en el sonido vocal debido al cierre irregular o asimétrico de los pliegues vocales. El ruido perjudica la claridad del sonido vocal y es percibido como ronquera. La HNR cuantifica la cantidad relativa de ruido en la voz [82]. Un mal cierre de los pliegues vocales aumenta la cantidad de ruido aleatorio. Si una persona tiene algún tipo de problema al hacer vibrar los pliegues vocales, escapa una mayor cantidad de aire durante la vibración, creando un ruido turbulento [83]. Cuanto más bajo es el HNR, más ruido hay en la voz, HNR alto indica un bajo nivel de ronquera.

2.2.3.3. Perturbación de la frecuencia (jitter)

El jitter es un índice que sirve para medir la regularidad en la frecuencia de la onda vocal; unas ondas puedan ser más anchas que otras. Los pliegues vocales tienen que vibrar a la misma velocidad para que la onda salga armónica. Naturalmente, como la voz humana no es una máquina, va a existir cierta variación natural. Los valores de referencia varían dependiendo con el programa que se mida. El jitter es uno de los factores que se ve afectado principalmente por el retardo del control de la vibración de las cuerdas vocales. Generalmente, en los pacientes que padecen alguna patología, el porcentaje de Jitter es más alto en comparación a un paciente sano [84].

2.2.3.4. Perturbación de la amplitud (Shimmer)

El shimmer mide irregularidades en la amplitud, el volumen o la intensidad, con la que se produce la onda: unas ondas pueden ser más alargadas que otras. Se relaciona con la emisión de ruido junto a la voz. Las medidas se toman ciclo a ciclo. Se estima que una voz es patológica a partir de un cierto porcentaje de irregularidades, el cual varía en la bibliografía según el programa con el que se mida [85].

2.2.3.5. Tiempo de inicio de la sonoridad (VOT)

EL VOT (voice Onset Time) es el tiempo transcurrido entre el final de una consonante y el inicio de la sonoridad vocálica. Requiere coordinación temporal entre la articulación oral y los mecanismos laríngeos requeridos para producir la vibración de las cuerdas vocales [86].Es el procedimiento habitual para cuantificar objetivamente la severidad de las disartrias.

2.2.3.6. Coeficientes Cepstrales de las frecuencias de Mel (MFCC)

Los Coeficientes Cepstrales de las frecuencias de Mel (MFCCs por sus siglas en inglés *Mel Frequency Cepstral Coefficients*) son coeficientes que permiten la representación del habla basada en la percepción auditiva humana [87]. Estos

muestran características locales de las señales de voz, asociadas al tracto vocal. Son ampliamente utilizadas en el área de verificación del interlocutor.

2.3. La Voz en el Parkinson

La EP se caracteriza por una pérdida progresiva de células mesencefálicas que se traduce en déficits motores que afectan los tres subsistemas implicados en el control del habla: el respiratorio, el fonatorio y el articulatorio. Este proceso conduce a un deterioro en la calidad del habla y la voz en aproximadamente el 90 % de los pacientes con EP. [88, 89, 90] Estas alteraciones se confunden en muchas ocasiones con los cambios naturales de los adultos mayores en relación con la presbifonía [91, 92], o estados depresivos [93].

Los trastornos vocales asociados con la enfermedad de Parkinson afectan la comunicación y la calidad de vida de los pacientes, independientemente de su edad o género [94].

El habla del parkinsoniano se distingue por tener una sonoridad e intensidad monótona, de bajo tono y pobremente prosódica, que tiende a desvanecerse al final de la fonación. El habla se produce en lentos ataques y significativas pausas para respirar entre palabras y sílabas, afectando la fluidez verbal [95] y el ritmo [96]. La articulación de los sonidos, tanto linguales como labiales, se empobrecen [97, 98], reduciendo significativamente su inteligibilidad [99] y dificultando la identificación de su estado emocional e intenciones [100].

Los pacientes presentan una disminución en la capacidad de los músculos laríngeos para mantener una posición fija al pronunciar de manera sostenida las vocales [101]. Varias investigaciones han evidenciado que estos pacientes tienen un tiempo de pronunciación de las vocales significativamente menor que las personas que no presentan la patología [102, 103]. Sakar en [104], demostró que las vocales sostenidas tienen información discriminatoria suficiente para la clasificación de la EP mediante el uso de modelos de aprendizaje automático.

El estudio de la voz aplicado a esta problemática se aborda principalmente

desde dos enfoques. Un enfoque busca estudiar distintos algoritmos para extraer características de la voz, y, luego, utilizan dichas características como entrada de clasificadores [17]. El otro enfoque emplea el deep learning sin buscar descriptores, y trabaja en el dominio original de la información (señales o espectrogramas de la voz).

2.3.1. Análisis de Características de la Voz

Diferentes trabajos han propuesto el procesamiento de señales de voz para obtener parámetros acústicos como método objetivo y no invasivo para la detección de la EP [15].

En los últimos años, se han propuesto varios sistemas computarizados para identificar los síntomas tempranos de la EP, monitoreando los patrones de voz [105, 106, 107]. En la Universidad de Oxford, [108, 14], se definieron un conjunto de medidas de disfonía y compararon los resultados de usar cuatro algoritmos de selección de características y clasificación binaria. Diaz en [109] presenta el desarrollo de un simulador biomecánico para modelar las particularidades del aparato fonador y han evaluado diferentes propuestas de extracción de parámetros para detectar la existencia de patologías utilizando la señal de voz de los pacientes.

Montaña en [110], propone diseñar un sistema experto para la detección temprana de la EP, a través del análisis articulatorio de la repetición rápida de sílabas como /pa-ta-ka/, calculando características temporales y espectrales extraídas en los segmentos de tiempo de inicio de voz.

Entre los hallazgos más frecuentes en la valoración de la voz de pacientes con EP se encuentra el incremento u otra variación de la F0, la reducción del tiempo de producción de vocales, el aumento del VOT, la disminución de la intensidad en la fonación, así como el decrecimiento del MPT (tiempo máximo de fonación), la perturbación del tono Jitter e intensidad Shimmer, y la razón ruido/armónicos, que también ha mostrado diferencias significativas [111]. Metter

y Hanson pusieron de manifiesto que el incremento en la F0 es paralelo a la gravedad de los síntomas y al avance de la enfermedad [103].

La alteración en prosodia expresiva está documentada y justificada por la reducción [112] variabilidad [103] e intensidad [113, 17], de la F0 en tareas de lectura de párrafos en los que los sujetos debían imitar frases acentuando su contenido emocional. Estos resultados ayudan en la búsqueda de los efectos que provoca esta enfermedad en la respiración, fonación, articulación y prosodia [114].

2.3.2. Clasificación Con Algoritmos de Aprendizaje Profundo

En vista de que, en la actualidad, no existen pruebas médicas disponibles que permitan un diagnóstico concluyente de la EP, los sistemas de diagnóstico asistido por computadora se presentan como una solución efectiva para la toma de decisiones basadas en datos y el apoyo al médico, especialmente en las primeras etapas de la enfermedad. Por otro lado, la extracción manual de características y la identificación de aquellas más relevantes para la detección de la EP requiere una gran cantidad de tiempo y recursos.

Una aproximación reciente para abordar este problema consiste en el estudio de la señal de audio y/o la representación visual del espectro de frecuencias de las señales de voz (espectrogramas) a través de técnicas de aprendizaje profundo. Esto permite que los datos se introduzcan directamente en la red, para que esta detecte las características y que, en base a ellas, realice la clasificación de manera automática.

En el año 2021, Alzubaidi [21] presentó un estudio cuyo objetivo fue el de explorar y resumir los sistemas de diagnóstico temprano asistidos por computadora para la EP. El estudio se llevó a cabo mediante la revisión de 91 trabajos seleccionados, en los cuales se encontró que, aproximadamente, la mitad de ellos habían implementado redes neuronales artificiales para realizar el diagnóstico de la enfermedad. Los conjuntos de datos biomédicos de voz y señales

fueron los tipos de datos más utilizados para el desarrollo y validación de estos modelos. Los trabajos analizados correspondieron al periodo comprendido entre los años 2018 y 2021.

Alhussein [107], presenta un sistema de detección de patologías de la voz basado en aprendizaje profundo mediante transferencia de aprendizaje sobre los modelos de CNN VGG16 [115] y CaffeNet [116] ajustados utilizando un clasificador de máquinas de vectores de soporte (SVM) [117]. La entrada de la red fue la representación visual del espectro de frecuencias de señales de voz capturadas mediante dispositivos móviles inteligentes. El sistema logró una precisión del 98,77 % en la base de datos Saarbruecken Voice Database (SVD) [118], con el modelo CNN de CaffeNet seguido por el clasificador SVM.

Vásquez-Correa en [119], presentan una metodología para clasificar EP a partir de muestras de audio en tres idiomas diferentes: español, alemán y checo. Se comparan dos enfoques, uno donde se extraen características a partir de las señales de voz, que se clasifican utilizando un SVM, y otro que utiliza espectrogramas para entrenar una CNN en un idioma, y luego transferir el aprendizaje a cada uno de los idiomas restantes. Los mejores resultados los obtuvieron al entrenar con el idioma español, alcanzando, una vez realizadas las transferencias, una precisión del 77.3 % en aleman y del 76.7 % en checo.

En Wodzinski [120], se utilizan para la detección de le EP una arquitectura ResNet [6] previamente entrenada usando las bases de datos ImageNet y SVD. Para la clasificación, utiliza los espectrogramas de audios de las vocales con fonación sostenida de la base PC-GITA [88]. La precisión obtenida en el conjunto de validación es superior al 90 %.

Zahid [121], en este trabajo, combina metodologías de aprendizaje profundo con metodologías de aprendizaje automático (SVM, random forest, perceptrón multicapa) para clasificar a los pacientes de Parkinson. Presenta la comparación de tres métodos para clasificar EP utilizando la base de datos PC-GITA. El primer método se basa en aprendizaje por transferencia aplicado a espectrogra-

mas de grabaciones de voz. El segundo método evalúa características profundas extraídas de espectrogramas de voz usando clasificadores de aprendizaje automático. El tercer método evalúa características acústicas simples de grabaciones usando clasificadores de aprendizaje automático. El enfoque basado en el segundo método dio la precisión más alta alcanzando un 99,7 % aplicando un perceptrón multicapa a grabaciones de monólogos.

Trinh y otros en [122], proponen un enfoque basado en redes neuronales de convolución para detectar patologías de la voz de personas con enfermedades neurodegenerativas como Parkinson y Alzheimer. Los experimentos se realizaron utilizando los espectrogramas de las muestras de voz de las bases de datos SVD y PC-Gita. Lograron una precisión de clasificación de más del 95 % sobre ambas bases.

Hireš y otros en [22], proponen un método de ajuste fino múltiple para entrenar una CNN. Utilizan las redes ResNet50 [6] y Xception [123] previamente entrenadas en el conjunto de datos ImageNet Para hacer el ajuste de la red trabajan con los espectrograma de grabaciones de voz mientras se pronuncian las vocales. Utilizan tres conjuntos de datos: el PC-GITA, la base de datos de voz SVD y el conjunto de datos de Vowels [124]. Después de entrenar el modelo usando SVD, Vowels y una subsección de PC-GITA, el modelo se valida usando el resto del conjunto de datos de PC-GITA. El rendimiento se midió utilizando la precisión, la sensibilidad, la especificidad y el área bajo la curva ROC (AUC). Si bien hubo pequeñas diferencias entre las diferentes vocales, el mejor desempeño fue cuando se consideró la vocal “a” sostenida (/a/); lográndose una precisión del 99 %, una sensibilidad del 86,2 %, una especificidad del 93,3 % y un AUC del 89,6 %.

Johri y otros en [125], estudian dos modelos basados en redes neuronales artificiales. Por un lado, un detector de espectrogramas obtenidos de las señales de marcha y un clasificador del deterioro de la voz. La precisión de clasificación en el detector de espectrogramas de marcha registra en un 88,1 %, mientras que el clasificador de deterioro de la voz ha mostrado una precisión del 89,15 %.

2.4. Fundamentos de los modelos de Deep Learning

El Deep Learning (DL), es una rama de la inteligencia artificial, que utiliza redes neuronales artificiales profundas para aprender de manera automática características complejas de los datos. A continuación, presentamos los conceptos clave para comprender cómo las máquinas pueden aprender y tomar decisiones basadas en datos.

2.4.1. Inteligencia artificial

Las definiciones de inteligencia artificial (IA) pueden variar, pero, en este contexto, puede entenderse como "la subdisciplina de la informática que busca desarrollar algoritmos y sistemas capaces de automatizar tareas intelectuales normalmente realizadas por humanos".

La IA ha experimentado una evolución significativa desde sus inicios en la década de 1950. Al principio, se basaba en un enfoque simbólico y se centraba en crear reglas explícitas para realizar tareas específicas. En este abordaje, los programadores escriben reglas para que la máquina, al procesar los datos, tome decisiones (en base a las reglas ingresadas) y resuelva problemas.

Con el avance de la tecnología y la disponibilidad de grandes cantidades de datos, la IA ha evolucionado hacia un enfoque conexionista donde se espera que las relaciones entre los datos surjan de forma automática, a través de la cooperación de un conjunto de elementos simples que forman parte de una red.

En la Figura 2 se puede ver que el campo de la IA engloba a las técnicas de Machine Leaning, mientras que el Deep Learning es un subcampo del Aprendizaje Automático.

2.4.2. Aprendizaje automático (Machine Learning)

Las técnicas de Machine Learning (ML), también conocido como aprendizaje automático, son un conjunto de herramientas y algoritmos diseñados para

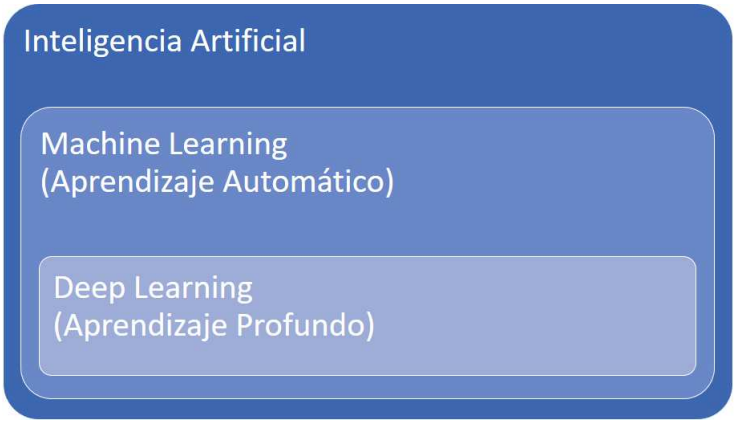


Figura 2: Inteligencia Articial incluye Machine Learning, que incluye Deep Learning

permitir que las máquinas aprendan y realicen tareas sin estar explícitamente programadas para cada tarea en particular. Estos sistemas se entrenan con grandes cantidades de datos para que puedan aprender por sí mismos cómo realizar tareas específicas.

En la programación clásica, a partir de un conjunto de reglas y datos, se obtienen resultados. Por ejemplo, si tenemos el dato: “3” y la regla: “sumar dos al dato”, el resultado que producirá el sistema será: “5”. Por otro lado, en un sistema de ML las entradas consisten en datos y las respuestas esperadas (etiquetas). El sistema de ML descubre la regla que permite que las entradas produzcan las respuestas esperadas. Siguiendo el ejemplo anterior, las entradas del sistema de ML serían: 1) el dato: “3” y 2) la respuesta esperada: "5"(etiqueta). El sistema proporcionará la regla: "sumar dos al datoçomo respuesta. Consultar la Figura 3 para más detalles.

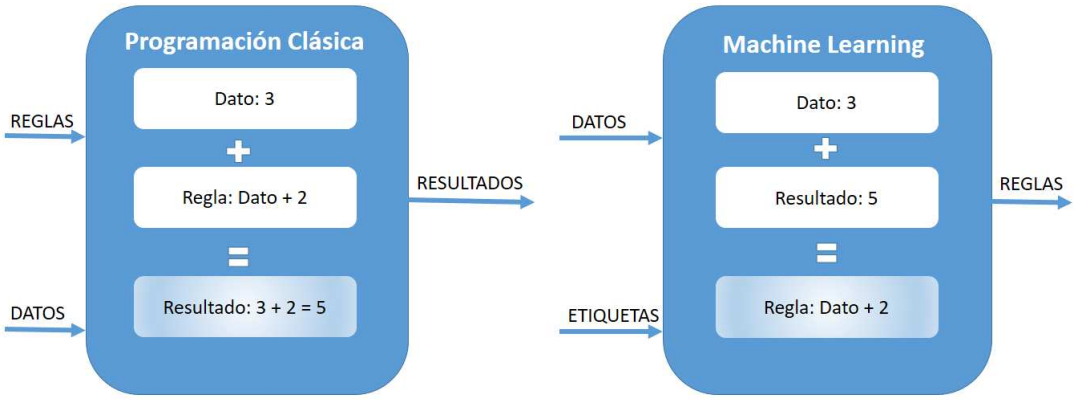


Figura 3: Programación Clásica vs técnicas de Machine Learning

Las técnicas de ML se utilizan para desarrollar sistemas de IA que pueden realizar una variedad de tareas, como reconocimiento de patrones, clasificación, predicción, toma de decisiones y optimización, entre otros. Estas se basan en el análisis de datos y la identificación de patrones a través de algoritmos estadísticos y matemáticos.

Los métodos tradicionales de ML requieren la intervención humana para determinar, de forma manual, el conjunto de características relevantes que el software debe analizar. Esto limita la capacidad del software, ya que depende de la habilidad del experto en los datos para detectar aquellas características más relevantes. Con el tiempo, estos sistemas pueden mejorar su desempeño a medida que reciben más datos y se les brinda más experiencia.

Como se puede ver en la Figura 4, la IA simbólica se enfoca en la programación de reglas y algoritmos para que una máquina realice tareas específicas, mientras que el ML se enfoca en el aprendizaje automático a partir de la experiencia y la retroalimentación.

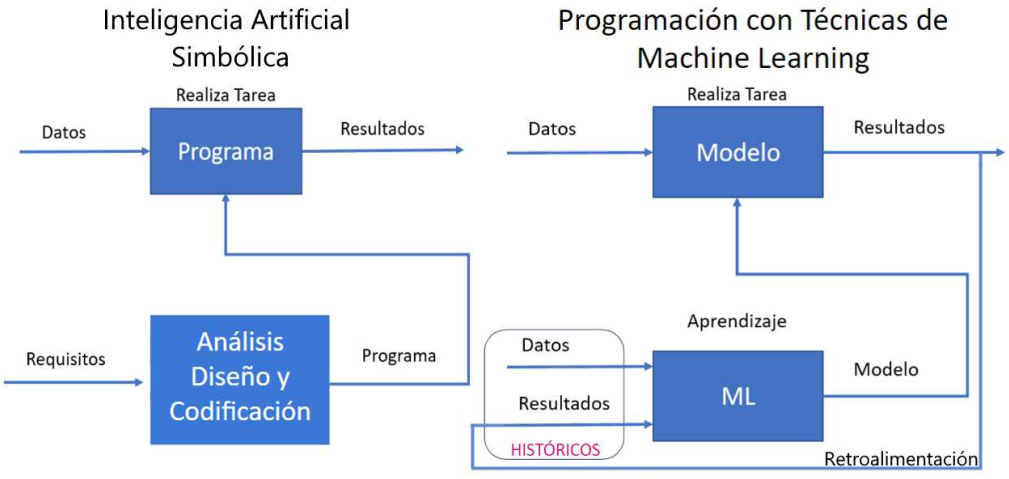


Figura 4: IA simbólica vs técnicas de aprendizaje automático

2.4.3. Aprendizaje Profundo o Deep Learning

El aprendizaje profundo es una rama del aprendizaje automático, que comparte los mismos conceptos fundamentales, pero difiere en los algoritmos utilizados. A diferencia del ML, donde se necesita la intervención de un experto

para seleccionar las características relevantes de los datos, en DL el científico de datos proporciona datos sin procesar a la red. La red de aprendizaje profundo obtiene y aprende las características de forma autónoma, ver Figura 5.

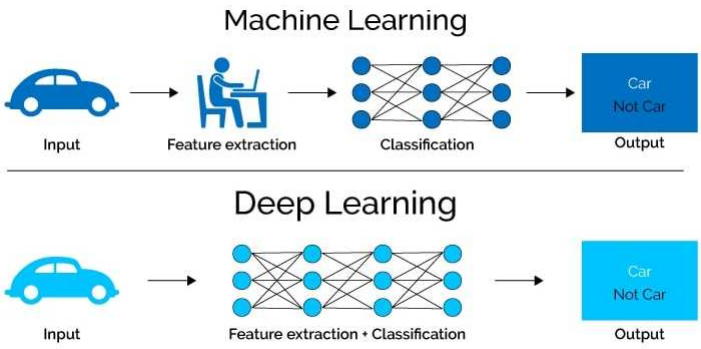


Figura 5: Técnicas de Machine Learning vs Deep Learning. Fuente [2]

El DL es particularmente bueno para la clasificación de imágenes, la generación de texto, la traducción automática y el reconocimiento de voz, entre las aplicaciones más sobresalientes.

2.4.4. Aprendizaje de un sistema de Machine Learning

El aprendizaje se entiende como la capacidad del sistema para identificar, extraer y reconocer patrones presentes en los datos. Se ingresan datos que, en este caso, son las características seleccionadas por un experto, así como las respuestas esperadas para esos datos (etiquetas) y se obtienen las reglas o modelo. Un modelo de ML se compone de una o varias funciones matemáticas que durante el entrenamiento ajustan sus parámetros para realizar una tarea específica. Luego, se aplican estos modelos a nuevos conjuntos de datos para producir respuestas originales. Un sistema de aprendizaje automático se entrena en lugar de programarse explícitamente.

2.4.5. Tipos de aprendizaje

Existen principalmente tres tipos de aprendizaje en ML: aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.

- Aprendizaje supervisado: el algoritmo recibe un conjunto de datos etiquetados, es decir, un conjunto de objetos con atributos y una etiqueta o categoría asociada a cada objeto. El objetivo del algoritmo es "aprender" o crear un modelo que permita predecir la etiqueta asociada a objetos desconocidos. La Figura 6 ilustra este tipo de aprendizaje. Este tipo de aprendizaje se utiliza comúnmente en tareas de clasificación y regresión.

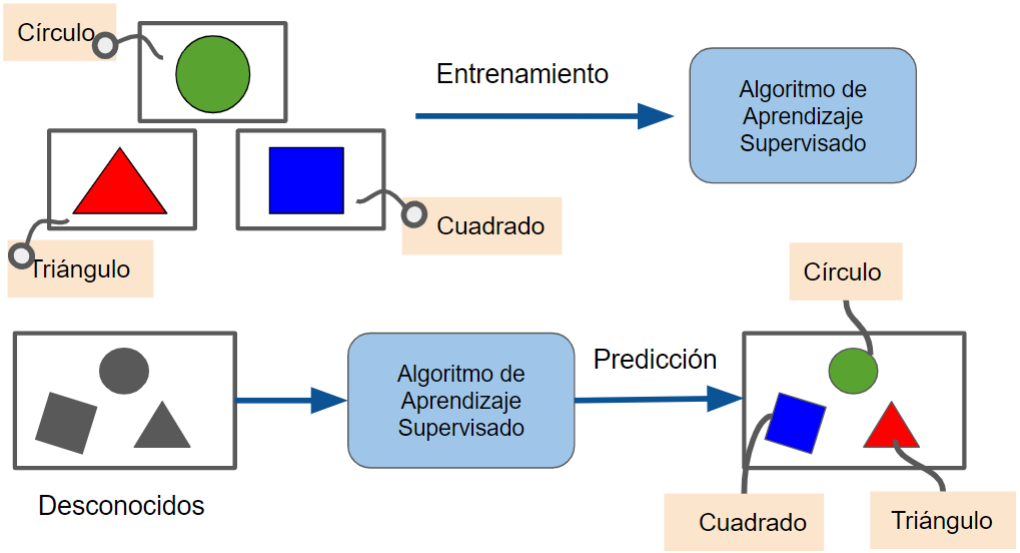


Figura 6: Aprendizaje supervisado

- Aprendizaje no supervisado: el algoritmo recibe un conjunto de datos sin etiquetar. El objetivo del algoritmo es aprender un modelo que permita descubrir patrones o estructuras en los datos. Por ejemplo, agrupar objetos similares en clusters, ver Figura 7. Este tipo de aprendizaje se utiliza comúnmente en tareas de reducción de dimensionalidad y clustering.

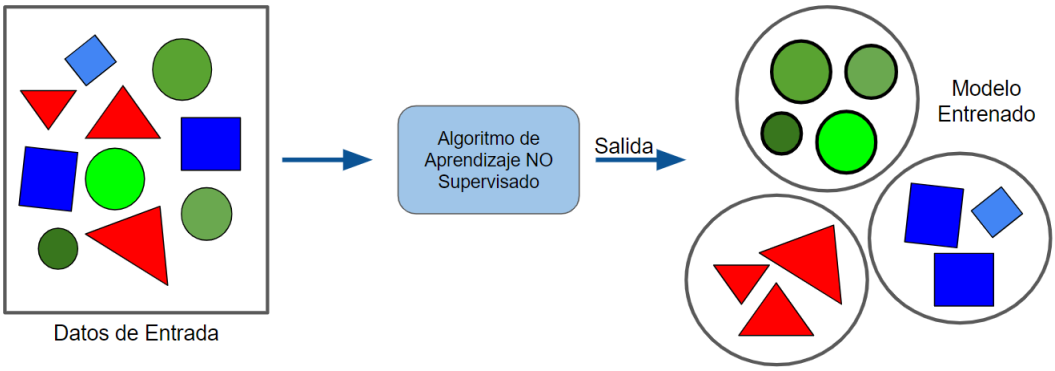


Figura 7: Aprendizaje no supervisado

- Aprendizaje por refuerzo: el algoritmo interactúa con un entorno y recibe recompensas o penalizaciones por sus acciones. El objetivo es aprender una política que maximice la recompensa total a lo largo del tiempo. Ver Figura 8. Este tipo de aprendizaje se utiliza comúnmente en tareas de control de sistemas y juegos.

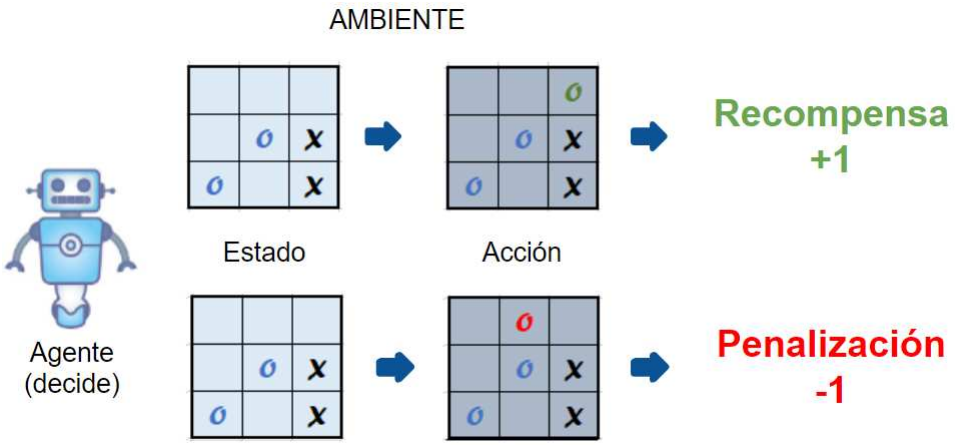


Figura 8: Aprendizaje por refuerzo.

2.4.6. Redes Neuronales Artificiales

Una red neuronal artificial es un modelo computacional inspirado en el funcionamiento del cerebro humano (Figura 9), ya que este órgano es capaz de procesar a gran velocidad grandes cantidades de información procedentes de los sentidos, combinarla o compararla con la información almacenada y dar respuestas adecuadas.

Figura 9: El modelo computacional de una red neuronal artificial se inspira en el funcionamiento del cerebro biológico.

Estos modelos se han basado sobre los estudios de las características esenciales de las neuronas y sus conexiones. Las neuronas están compuestas de dendritas, soma (cuerpo), núcleo y el axón. Las dendritas se encargan de captar los impulsos nerviosos que emiten otras neuronas. Estos impulsos se procesan en el soma y se transmiten a través del axón que emite un impulso nervioso hacia las neuronas contiguas.

Las neuronas artificiales imitan el comportamiento de la neurona biológica, ver Figura 10. Las conexiones que llegan a una neurona, tiene valores asociados (pesos) que definen con qué intensidad cada variable de entrada afecta a la neurona y produce una salida.

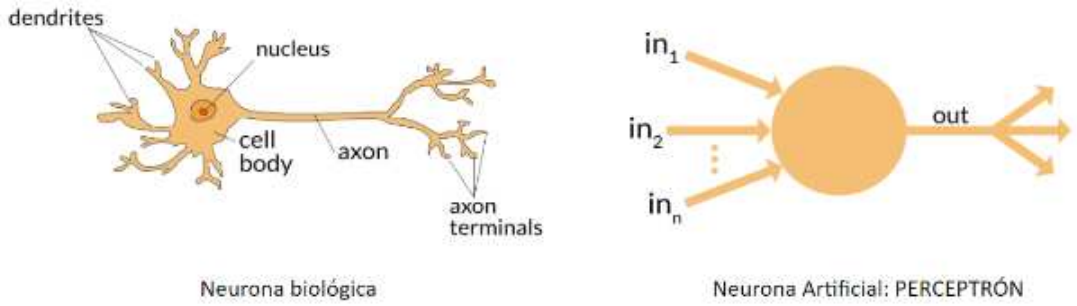


Figura 10: Comparación entre una neurona biológica y una neurona artificial.

Al igual que una neurona biológica, una neurona artificial recibe entradas de las capas anteriores a través de las conexiones. Como se ilustra en la Figura 11, una neurona "i" recibe las entradas $x_1, x_2, \ldots x_j \ldots x_n$ de las capas anteriores (dendritas). Cada conexión tiene un peso ponderado $w_{i1}, w_{i2}, \ldots w_{ij}, \ldots w_{in}$. La neurona (cuerpo celular) procesa la entrada y produce una salida basada en su función de activación. Cada entrada "x_j" se multiplica por el peso ponderado de su conexión "w_{ij}" y se suma a otros pesos de entrada. A esta suma ponderada se aplica la función de activación "$f()$"para producir la salida "y_i" (axón). La función de activación puede ser una función lineal o no lineal, dependiendo de la tarea que se esté abordando. El "bias"(umbral) es un parámetro que se utiliza para ajustar la salida de la neurona.

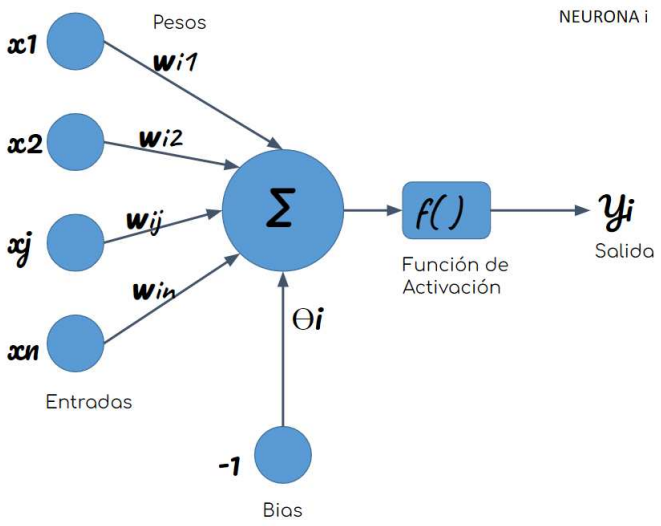


Figura 11: Esquema general de una neurona artificial.

2.4.6.1. Modelo de una Neurona Artificial

La descripción de una neurona artificial, siguiendo el "modelo estándar de neurona artificial" descritos en [126, 127], para la neurona i-ésima será:

- Dado un conjunto de entradas x_j y unos pesos sinápticos w_{ij}, donde los valores de j = 1,..., n.
- Una regla de propagación h_i definida a partir del conjunto de entradas y los pesos sinápticos. Es decir:

  $$\text{Regla de Propagación} = h_i(x_1, \ldots, x_n, w_{i1}, \ldots, w_{in})$$

  La regla de propagación más habitual consiste en combinar linealmente las entradas y los pesos sinápticos, obteniéndose:

  $$h_i(x_1,...,x_n,w_{i1},...,w_{in}) = \sum_{j=1}^{n} w_{ij}x_j \tag{1}$$

  Cada neurona tiene un parámetro adicional θi, que se denomina umbral o bias. El objetivo del bias es permitir que la neurona tenga cierta flexibilidad en la salida para ajustarse mejor a los datos de entrada. Agregando el bias:

  $$h_i(x_1,...,x_n,w_{i1},...,w_{in}) = \sum_{j=1}^{n} w_{ij}x_j - \theta_i) \tag{2}$$

- Una función de activación que representa simultáneamente la salida de la neurona y su estado de activación. Si se denota por y_i a la salida de la neurona, esto equivale a aplicar la función de activación f_i al conjunto de entrada de dicha neurona h_i, así se obtiene:

$$y_i = f_i(h_i) = f_i(\sum_{j=1}^{n} w_{ij}x_j - \theta_i) \tag{3}$$

2.4.6.2. Red Neuronal Artificial

Una red neuronal artificial se compone de un conjunto de nodos (neuronas o unidades de procesamiento simple), que se organizan en capas. Las capas intentan imitar el proceso de sinapsis de las neuronas. Estas conexiones se utilizan para propagar la información desde la entrada hasta la salida del modelo. Las mismas se pueden ajustar mediante el proceso de aprendizaje. Es decir, se ajustan a partir de la experiencia (datos). La red está formada por una capa de entrada por donde ingresan los datos, una capa de salida donde la red hace su predicción y puede tener una o varias capas ocultas, como se muestra en la Figura 12, que son las que permiten aprender características más complejas de los datos de entrada. La profundidad de la red se refiere a la cantidad de capas ocultas que la forman.

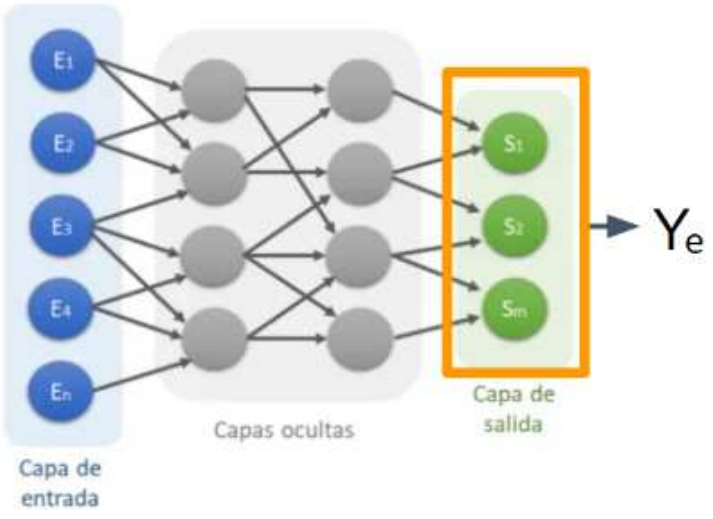


Figura 12: Esquema de una red neuronal artificial.

Cada capa de la red realiza una transformación de los datos de entrada. Las características más complejas se aprenden a través de la combinación lineal y no lineal de las características simples aprendidas en capas previas. A medida que los datos se procesan a través de las capas de la red, se van aprendiendo patrones más complejos y abstractos en los datos de entrada, lo que a su vez permite la resolución de problemas más complejos.

El proceso de aprendizaje de la red consiste en ajustar los pesos asociados a cada conexión entre las neuronas, con el objetivo de minimizar el error entre la salida predicha por el modelo y las etiquetas reales en el conjunto de entrenamiento. Este proceso se realiza iterativamente hasta que el modelo alcance una precisión deseada. Una vez entrenado, el modelo puede ser utilizado para hacer predicciones sobre nuevos datos, que nunca ha visto antes.

2.4.7. Arquitecturas de Redes Neuronales

La arquitectura de una red neuronal hace referencia a la forma en que están organizadas las capas y los nodos (neuronas) dentro la red. Incluye el número de capas, el número de nodos en cada capa, y cómo están conectados los nodos entre capas. La elección de la arquitectura adecuada para un problema depende de muchos factores, como la cantidad y el tipo de datos de entrada, la complejidad del problema, requisitos de tiempo de ejecución y memoria.

Algunas de las arquitecturas más representativas de las redes neuronales son:

- Redes neuronales Feedforward: son las redes más simples, donde los datos fluyen en una dirección desde la entrada hasta la salida, sin ciclos o retroalimentación. Las salidas de una capa se utilizan como entrada en la capa siguiente, pero no hay retroalimentación de las salidas hacia las capas anteriores.
    - Perceptrón: Se compone de una capa de entrada, una capa oculta y una capa de salida.
    - Perceptrón Multicapa (MLP): Es una extensión del perceptrón que incluye más de una capa oculta.

- Redes neuronales convolucionales (CNN): Es una arquitectura de redes neuronales diseñada específicamente para trabajar con imágenes. Utiliza filtros para extraer características importantes de los datos de entrada.
- Redes neuronales recurrentes (RNN): Es una arquitectura de redes neuronales en las que los datos pueden retroalimentarse y fluir en ambas direcciones. Está diseñada para trabajar con secuencias de datos, como texto o audio.
- Autoencoder: Es una arquitectura de redes neuronales que aprende una representación compacta de los datos de entrada. Se compone de un encoder y un decoder, donde el encoder codifica los datos de entrada en una representación de baja dimensionalidad y el decoder los reconstruye.

2.4.8. Red Neuronal Convolucional (CNN)

Las redes neuronales convolucionales (*Convolutional neural network*, también llamadas ConvNets), son un tipo de arquitectura de DL, específicamente diseñadas para aprender y descubrir características a partir de imágenes [128, 129]. Son un modelo donde las neuronas corresponden a campos receptivos de una manera similar a las neuronas de la corteza visual primaria de un cerebro biológico [130]. Se han utilizado con éxito en una amplia variedad de tareas relacionadas con imágenes, como la clasificación de imágenes, la detección de objetos y la segmentación de imágenes, procesamiento del lenguaje natural, procesamiento de audio, entre otros [131]. Los algoritmos de aprendizaje profundo buscan en el aprendizaje automático características de alto de alto nivel a partir de grandes volúmenes de datos [132].

2.4.9. Arquitectura típica de las CNNs

La arquitectura de una CNN típica se ilustra en la Figura 13. Esta arquitectura típica puede variar dependiendo de la tarea específica y de las necesidades de la aplicación. Está estructurada como una serie de capas divididas en dos etapas

básicas. Las capas de la primera etapa se encargan de extraer las características de la imagen de entrada. Las capas de la segunda etapa se encargan de la clasificación de la imagen asignándole una etiqueta o clase. A continuación, se describen las capas más comunes de una CNN:

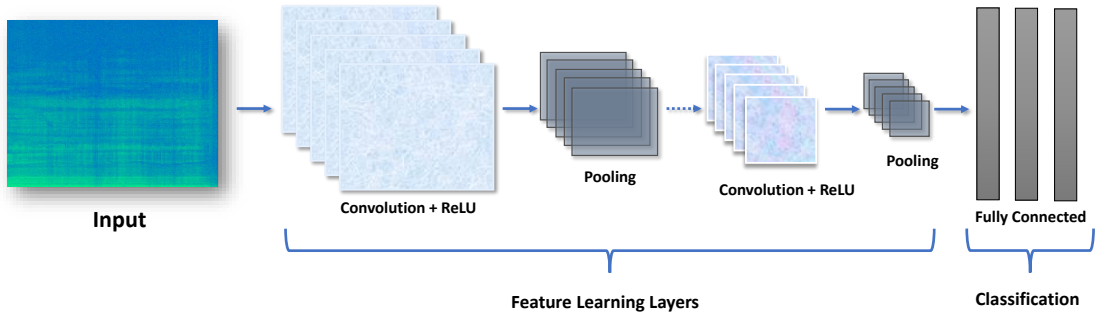


Figura 13: Arquitectura típica de una CNN.

- Capa de entrada: esta capa toma la imagen de entrada y la convierte en una representación numérica para su procesamiento por la red.
- Capa de Convolución: esta capa es el componente central de una CNN. Su función principal es extraer características y patrones de una imagen. Utiliza filtros (o kernels) que se desplazan por la imagen realizando cálculos elemento a elemento. De esta manera, se obtiene un mapa de características que representa las características relevantes de la imagen.
- Capa de ReLU (Rectified Linear Unit): esta capa agrega no linealidad a la salida de la capa de Convolución, lo que permite a la red modelar relaciones más complejas en los datos.
- Capa de pooling: esta capa reduce la dimensionalidad de la salida de la capa anterior (disminuye la cantidad de parámetros) , lo que disminuye los tiempos de procesado, y ayuda a la red a ser más robusta frente a pequeñas transformaciones en la entrada. La capa de pooling utiliza técnicas como el pooling máximo o el pooling promedio. La reducción de parámetros se realiza mediante la extracción de estadísticas, como el promedio o el máximo, de una región fija del mapa de características.

- Capas de Convolución y ReLU adicionales: dependiendo de la complejidad de la tarea, y del tamaño de la imagen de entrada, puede haber varias capas adicionales de Convolución y ReLU en la arquitectura de la CNN.
- Capa fully connected o Capa totalmente conectada (Dense Layer): Esta capa se denomina densa, ya que todos los nodos están conectados entre sí, y aplana la salida de las capas anteriores a una dimensión. Combina todas las características extraídas por las capas previas, y las usa para producir una predicción.
- Capa de salida: esta capa produce la salida final de la CNN, que puede ser una probabilidad para cada clase en un problema de clasificación, o una etiqueta única para la imagen.

2.4.10. Breve descripción de las Arquitectura de CNNs

En esta tesis se utilizaron 5 arquitecturas de redes CNN para realizar los diferentes experimentos. Las redes utilizadas fueron AlexNet, VGG16, ResNet50, Inception V3 y Squeezenet. A continuación, se detalla una breve descripción de cada una de ellas:

- AlexNet [3]: Fue una de las primeras redes profundas. Tiene una arquitectura secuencial compuesta de 8 capas: 5 convolucionales y 3 capas totalmente conectadas.

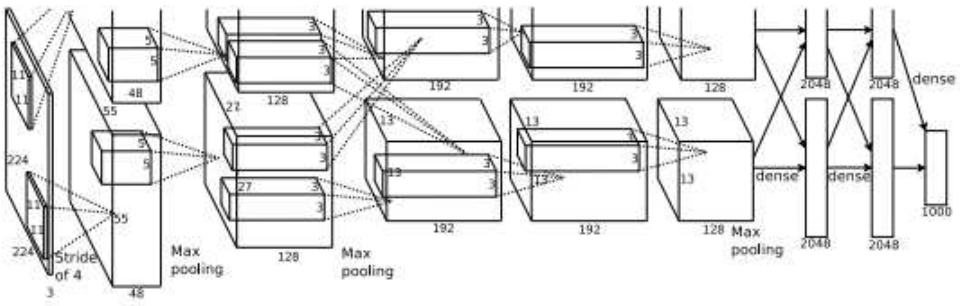


Figura 14: Esquema de la red AlexNex, publicado en el artículo de Krizhevsky. El mismo muestra explícitamente la delimitación de responsabilidades entre las dos GPU utilizadas durante su entrenamiento. Fuente [3]

- VGG-16 [115]: Es una mejora de AlexNet. La red VGG-16 fue desarrollada por el Grupo de Geometría Visual (VGG) de la Universidad de Oxford. Está formada por 13 capas convolucionales para la extracción de características y 3 capas densas. Al ser mas profunda permite aprender características más complejas.

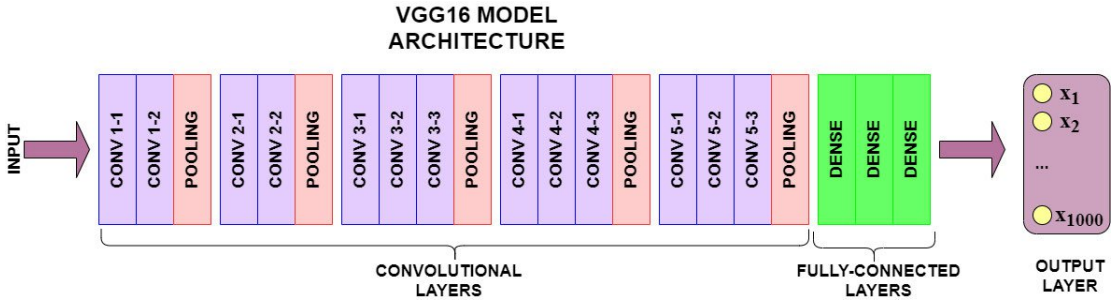


Figura 15: El modelo VGG16 tiene 13 capas convolucionales, 3 capas densas y una capa de salida de 1000 nodos. Fuente [4]

- Inception V3 [133]: Esta CNN pertenece a la familia de redes del tipo Inception. Para evitar el problema del sobreajuste que se produce al tener una red muy profunda, se introduce el concepto del procesamiento en paralelo de distintos filtros con múltiples tamaños dentro de un módulo. La salida de cada módulo es la concatenación de los resultados obtenidos. La estructura de la red se vuelve más amplia que profunda. Tiene un alto rendimiento de clasificación con un costo computacional menor que el de VGGNet.

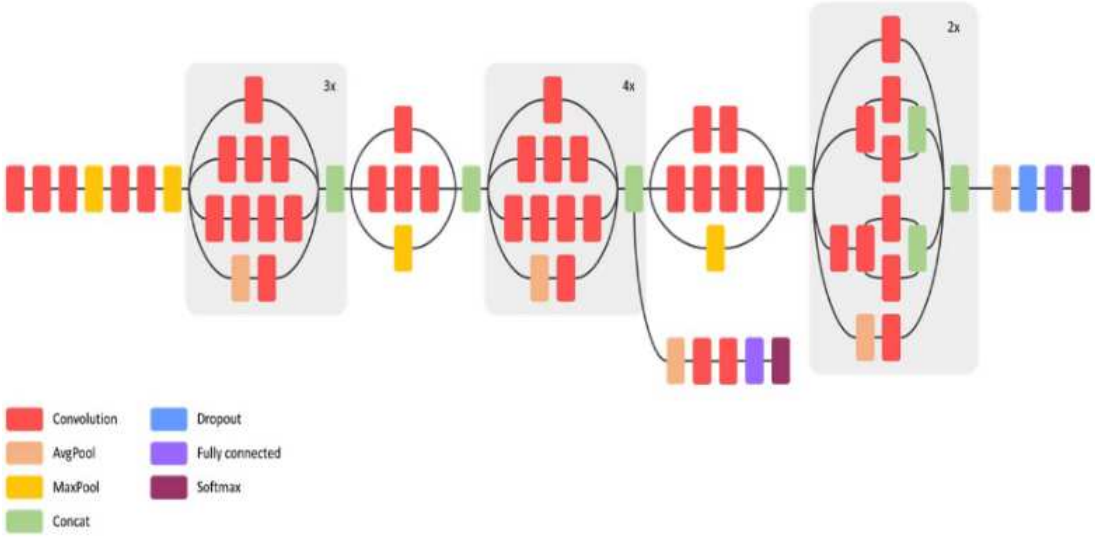


Figura 16: Esquema de la arquitectura de una red Inception V3. Fuente [5]

- ResNet (ResNet-50) [6]: Para evitar el problema del desvanecimiento del gradiente que se produce al seguir aumentando el número de capas, se introdujo la utilización de bloques residuales. Esta arquitectura permitió una reducción en el número de parámetros a estimar, manteniendo una buena relación entre el rendimiento de la red y tiempo de entrenamiento.

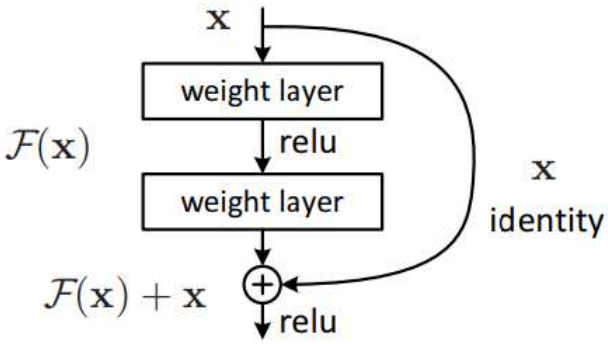


Figura 17: ResNet: un bloque de construcción del aprendizaje residual. Fuente [6]

- Squezeenet [7]: Es una red compacta diseñada para operar en equipos con pocos recursos y de fácil transmisión a través una red. Su arquitectura se basa en el uso de módulos "Fire", que constan de una capa de compresión y una capa de expansión. La tasa de aciertos alcanzada es similar a la de AlexNet en ImageNet, pero utilizando 50 veces menos parámetros, y funcionando 3 veces más rápido.

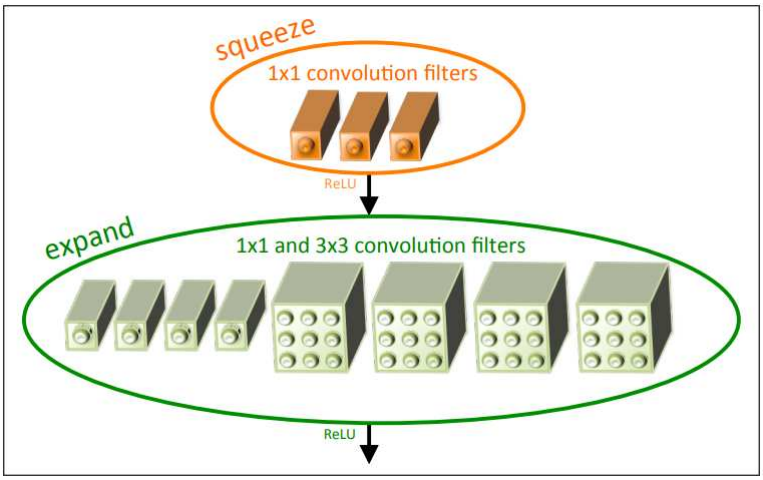


Figura 18: Vista de microarquitectura: Organización de filtros de convolución en el módulo Fire. En este ejemplo, s1x1 = 3, e1x1 = 4 y e3x3 = 4. Se ilustran los filtros de convolución pero no el de activaciones. Fuente [7]

En la Tabla 2.1 se presentan las principales característica de las cinco CCN utilizadas en esta tesis. Profundidad hace referencia a la cantidad de capas de la red, el tamaño indica la cantidad de mega bytes que ocupa la red, parámetros indica en millones la cantidad de parámetros de cada una y por último el tamaño en píxeles de las imágenes de entrada a cada red.

Tabla 2.1: Parámetros y dimensiones de las CNNs utilizadas en esta tesis [8]

Network	**AlexNet**	**VGG-16**	**SqueezeNet**	**Inception V3**	**ResNet-50**
Profundidad	8	16	18	48	50
Tamaño (MB)	227	515	4.6	89	96
Parámetros (Milliones)	61	138	1.24	23.9	25.6
Tamaño de la imagen de entrada	227×227	224×224	227×227	299×299	224×224

2.4.11. Aprendizaje por transferencia

El aprendizaje por transferencia es una técnica fundamental, en el campo del aprendizaje profundo, que permite a un modelo aplicar los conocimientos adquiridos en una tarea a otra tarea relacionada. En lugar de entrenar un modelo desde cero para cada tarea específica, esta técnica reutiliza el conocimiento previamente aprendido para mejorar su rendimiento en nuevas tareas. Se trata de reutilizar un modelo previamente entrenado sobre un conjunto de datos, y utilizarlo como punto de partida para aprender las características de nuevos conjuntos de datos que, por lo general, suelen ser más reducidos. Al aplicar el aprendizaje por transferencia, se reduce el tiempo de entrenamiento del modelo por utilizar los patrones aprendidos previamente, mejorandose significativamente su rendimiento, eficacia, y capacidad de generalización, especialmente en situaciones en las que los datos son escasos, o costosos de obtener.

2.4.11.1. Descripción del proceso de general de Aprendizaje por transferencia

Como se mencionó en el apartado anterior, la transferencia de aprendizaje es una técnica de aprendizaje profundo que permite aprovechar el conocimiento adquirido, por una red neuronal pre-entrenada, en una tarea para mejorar el

rendimiento de otra tarea relacionada. La Figura 19 ilustra el proceso de transferencia de aprendizaje sobre las CNN, el cual, generalmente, sigue los siguientes pasos:

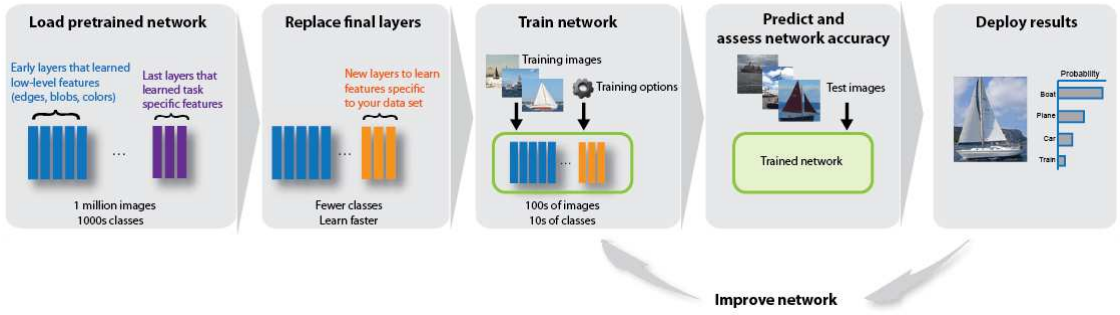


Figura 19: Proceso de transferencia de aprendizaje. Fuente [8]

- Cargar una red pre-entrenada: se selecciona una red pre-entrenada sobre un conjunto de datos de gran escala, como ImageNet.
- Reemplazar las capas finales: se eliminan las últimas capas de la red pre-entrenada, que son específicas de la tarea original, y se agregan nuevas capas que se adaptan a la nueva tarea que se desea resolver.
- Entrenar la red: se utiliza el conjunto de imágenes de entrenamiento de la nueva tarea para entrenar la red adaptada. Durante este proceso, las nuevas capas aprenden las características específicas de las imágenes del nuevo conjunto de datos, mientras que las capas pre-entrenadas ya han aprendido las características de bajo nivel, como bordes, manchas y colores.
- Evaluar la nueva red: se utiliza el conjunto de datos de prueba para evaluar la precisión de la red adaptada.
- Analizar los resultados: se analiza el desempeño de la nueva red y se ajustan los parámetros de entrenamiento si es necesario.

En resumen, la transferencia de aprendizaje es una técnica efectiva para mejorar la precisión de una red neuronal en una tarea específica, especialmente cuando el conjunto de datos de entrenamiento es pequeño o los recursos de computación son limitados.

2.4.12. Máquina de Aprendizaje Extremo

Las Máquinas de Aprendizaje Extremo (ELM) fueron propuestas inicialmente en el año 2006 por Huang en [27] como un algoritmo de entrenamiento rápido, y preciso, para redes neuronales feedforward de una sola capa oculta (*Single Layer Feedforward Neural Network - SLFN*). Sin embargo, ELM hace parte de la familia de *randomization - based feedforward neural networks* [134, 135]. Otros algoritmos que hacen parte de esta familia son *radial basis function (RBF) network* [136, 137], *feedforward neural networks with random weights (RWNN)* [138], *random vector functional link (RVFL) net* [139], y *recurrent neural networks* [140]. ELM tiene varias similitudes con estos algoritmos, especialmente con RWNN y RVFL, que fueron propuestas en 1992 y 1994 respectivamente. RWNN, es de los primeros algoritmos SLFN no iterativo que implementa aleatoriedad en su estructura. El algoritmo mantiene fijos los pesos de la capa oculta y asigna sesgos aleatorios [138]. Por su parte, RVFL incorpora un enlace funcional con enlaces directos desde la capa de entrada a la capa oculta, asignando aleatoriamente los pesos y sesgos de la capa oculta [139]. Esto implica que las entradas, para la capa de salida, es la suma de neuronas en la capa de entrada y la capa oculta. Al igual que ELM, RVFL tiene diferentes tipos de arquitectura, donde destaca *shallow RVFL, ensemble learning based RVFL, Deep RVFL, y ensemble Deep RVFL* [141]. Por estas similitudes, algunos autores consideran FWNN y RVFL la fuente de inspiración de ELM.

Las ELM se caracterizan por tener una alta velocidad de convergencia, una elevada capacidad de generalización, y por requerir un menor número menor de parámetros de entrenamiento [27, 142, 143, 144, 145]. A diferencia de las SLFN tradicionales, en la cuales se intenta encontrar los pesos y sesgos de toda la red mediante un algoritmo de optimización, en las ELM los parámetros de la capa oculta se determinan aleatoriamente, y los parámetros de la capa de salida se estiman mediante la resolución de un sistema de ecuaciones lineales sobredeterminado, el cual se resuelve de forma eficiente, y óptima, utilizando la pseudoinversa de Moore-Penrose [145].

En problemas de aprendizaje supervisado, la salida de la ELM viene dada por la siguiente ecuación:

$$f_L(\boldsymbol{x}_j) = \sum_{i=1}^{L} \boldsymbol{\beta}_i g(\boldsymbol{w}_i \boldsymbol{x}_j + b_i) = \boldsymbol{t}_j, \quad j = 1, \ldots N \tag{4}$$

Donde L representa el número de nodos ocultos, $w_i = [w_{i1}, w_{i2}, ..., w_{in}]^T$ la matriz de pesos que conecta el i-ésimo nodo oculto con el j-ésimo ejemplo de entrenamiento (x_j), b_i los sesgos de la capa oculta, g es una función de activación, $\beta_i = [\beta_{i1}, \beta_{i2}, ..., \beta_{im}]^T$ representa los pesos de la capa de salida, t_j las etiquetas de los x_j, y N el número de muestras. La ecuación (4) se puede escribir de forma compacta como sigue:

$$\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{T} \tag{5}$$

donde:

$$\boldsymbol{H} = \begin{bmatrix} g(\boldsymbol{w}_1 \boldsymbol{x}_1 + b_1) & \cdots & g(\boldsymbol{w}_L \boldsymbol{x}_1 + b_L) \\ \vdots & \ddots & \vdots \\ g(\boldsymbol{w}_1 \boldsymbol{x}_N + b_1) & \cdots & g(\boldsymbol{w}_L \boldsymbol{x}_N + b_L) \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1^T \\ \vdots \\ \boldsymbol{\beta}_N^T \end{bmatrix}, \boldsymbol{T} = \begin{bmatrix} \boldsymbol{t}_1^T \\ \vdots \\ \boldsymbol{t}_N^T \end{bmatrix} \tag{6}$$

H se corresponde a la matriz de salida de la capa oculta [27, 146, 147], T es la matriz de etiquetas del conjunto de datos, y β es la matriz de pesos de la capa de salida. La expresión para calcular β es la siguiente:

$$\boldsymbol{\beta} = \boldsymbol{H}^{\dagger}\boldsymbol{T} \tag{7}$$

siendo $\boldsymbol{H}^{\dagger}$ la inversa de Moore-Penrose de la matriz $\boldsymbol{H}$ [148].

Quedando definidas entradas y salidas como:

Entrada: Un conjunto de entrenamiento

$\aleph = \{(x_i, t_i) : x_i \in R^n, t_i \in R^m, i = 1, ..., N\}$, g y L.

Salida: β.

y el algoritmo del entrenamiento supervisado resumido en los siguientes 3 pasos [27]:

Paso 1: Asignar valores pseudo-aleatorios a (w_i) y (b_i);

Paso 2: Calcular $\boldsymbol{H}$ según la ecuación (6).

Paso 3: Calcular $\boldsymbol{\beta}$, según la ecuación (7).

2.4.13. Espectrogramas

Los sonidos son ondas, formados por una mezcla de diferente frecuencias o sonidos puros. En Figura 20 se pueden observar los elementos que la caracterizan una onda de sonido pura:

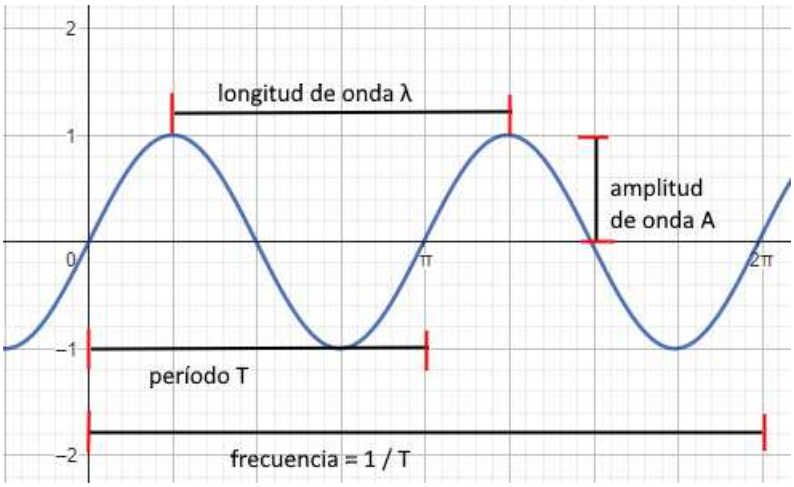


Figura 20: Elementos característicos de una onda de sonido.

- La longitud de onda: es la distancia mínima en la que se repite un ciclo de una onda.
- El Período: es la cantidad de tiempo que tarda en finalizar una revolución completa de un ciclo de onda.
- La amplitud: es la altura de onda y está relacionada con la intensidad del sonido (volumen).
- La Frecuencia: es el número de ondas o ciclos completos producidos en un segundo. Cuanto mayor es la frecuencia más agudo es el sonido.

Cuando un sonido es descompuesto en un eje de tres dimensiones: frecuencia, amplitud y tiempo, se pueden representar todos los elementos que lo componen, como se puede observar en la Figura 21. Sin embargo, al intentar visualizarlo en un gráfico bidimensional, inevitablemente se pierde información. En el dominio del tiempo, las ondas simples se suman y se vuelven indistinguibles. En el dominio frecuencial, es posible observar la amplitud correspondiente a cada frecuencia presente.

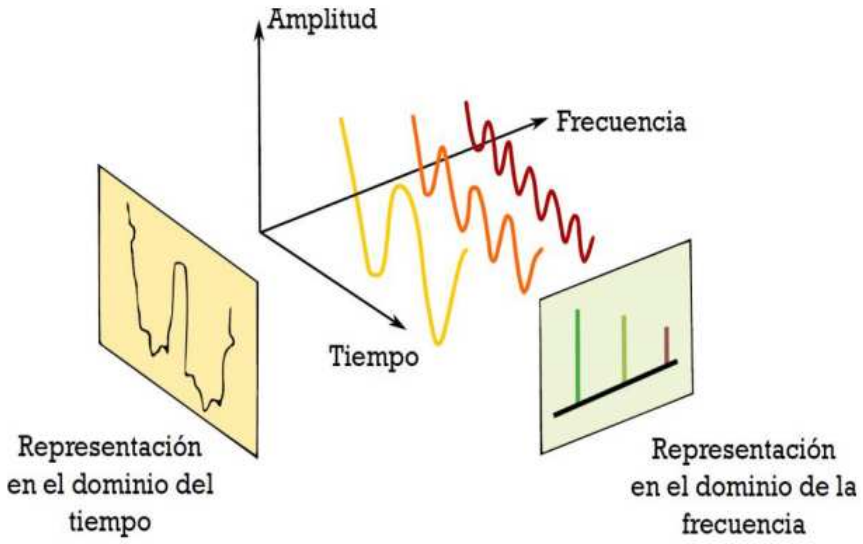


Figura 21: Representación de una onda de sonido en 3 dimensiones.

El espectrograma es una representación visual de una señal, que muestra cómo se distribuye la energía en función del tiempo, la frecuencia y la amplitud (ver Figura 22). A pesar de tener tres dimensiones, puede ser visualizado en una imagen bidimensional. En el espectrograma, las variaciones de color representan la intensidad del sonido a lo largo del tiempo (eje horizontal) en relación con la frecuencia (eje vertical). La representación del habla a través de espectrogramas ha demostrado ser estable y robusta aún con altos niveles de ruido [19, 20].

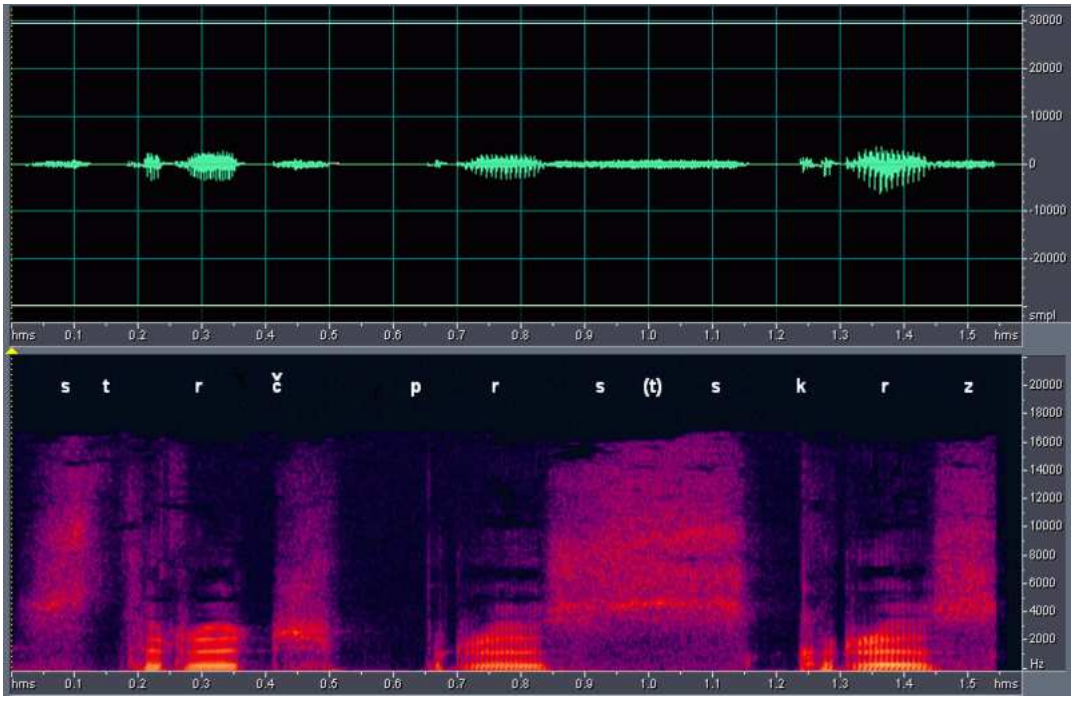


Figura 22: Espectrograma de la oración del checo Strč prst skrz krk que significa 'Introduce el dedo a través de la garganta'.

Para crear un espectrograma, se aplica la Transformada de Fourier de tiempo corto (*Short-Time Fourier Transform - STFT*), a la señal de audio. La STFT divide el audio en pequeñas ventanas de tiempo. Para cada una de estas ventanas se calcula la frecuencia y la amplitud. Para formar el espectrograma, cada ventana de tiempo es una columna del espectrograma donde, a cada combinación de tiempo y frecuencia, se le asigna un color según la amplitud. La Figura 23 ilustra un ejemplo donde la señal disminuye la frecuencia en cada ventana de tiempo, al mantener constante su amplitud a lo largo de todo el período se le asigna el color rojo uniforme.

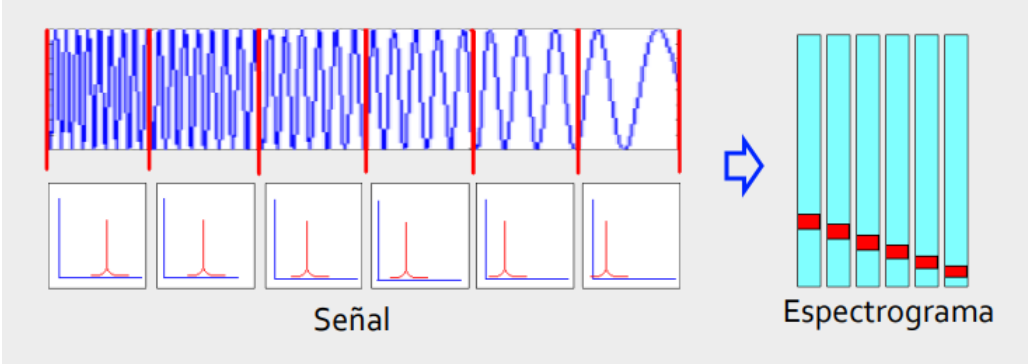


Figura 23: Ejemplo de aplicación de la STFT a una señal de sonido con amplitud constante.

Página intencionalmente en blanco

Capítulo 3

Metodología de los experimentos

3.1. Hardware y Software

Para llevar a cabo la implementación de los códigos se utilizó el Lenguaje MatLab R2018b. El Image Processing Toolbox fue utilizado para el procesamiento de las imágenes. El Deep Learning Toolbox para el entrenamiento de las Redes Neuronales Convolucionales, donde se instalaron los modelos de AlexNet, ResNet-50, VGG-16, SqueezeNet e Inception-v3. Para el entrenamiento de las Redes de Aprendizaje Extremo se utilizaron las rutinas del Toolbox del sitio web oficial de Extreme Learning Machine.

3.2. Base de Datos de Audios

La base de datos (BD) de voces utilizadas en este trabajo [149] fue desarrollada durante el año 2019 por un grupo interdisciplinario coordinado por investigadores de la Universidad Nacional de La Matanza (UNLaM), sumado a personal de la salud del Hospital Nacional Alejandro Posadas, y del Hospital Bernardino Rivadavia, con el asesoramiento del Dr. Jorge Gurlekian.

Este repositorio contiene datos de 55 pacientes con EP, 24 mujeres y 31 varones, la media de edades es de 64 años (las mismas varían entre los 38 y 79 años), y datos de 64 personas sin EP del mismo rango etario, de las cuales el 68 % son mujeres y el 32 % varones. Se grabaron entre 1 a 3 muestras por cada

persona. Los enfermos de Parkinson fueron evaluados neurológicamente con la escala Hoehn & Yarh y el test UPDRS en español en la versión patrocinada por la Sociedad de Trastornos del Movimiento (Movement Disorders Society, MDS)[150]. Además, se les realizó una endoscopía de las cuerdas vocales. El tiempo de evolución promedio de la enfermedad desde el diagnóstico es de 6 años con valores que van entre los 6 meses y los 16 años. En el momento de realizar las grabaciones, los enfermos de Parkinson se encontraban medicados. Es decir en, "ON", estado que indica que los síntomas son de baja intensidad.

La base de enfermos y sanos contiene muestras de habla espontánea, la fonación de las vocales sostenidas /a/, /i/, /u/ por separado y por un tiempo estimado de 3 a 5 segundos. Además, se incluyó la repetición de la palabra "Pataka" y la fonación de la frase: "¡Betty! ¡Qué inmensa alegría escucharte! Cuando vengas para fin de año, quiero llevarte a recorrer toda la Argentina." Las voces fueron grabadas por un técnico de sonido en un espacio especialmente acondicionado, con un micrófono condensador polarizado permanente de placa trasera con carga fija (AT2020 micrófono de condensador cardioide). Los audios están grabados en una frecuencia de muestreo de 44.100 Hz.

La base de datos con voces de personas con y sin EP estará disponible en el repositorio de la UNLaM (https://repositoriocyt.unlam.edu.ar/) con acceso público.

3.3. Base de datos de espectrogramas

Para la construcción del repositorio de espectrogramas se utilizó la base de datos presentada en el apartado anterior.

Se trabajó sobre los sonidos originales, es decir, aquellos cuya duración corresponde al tiempo que la persona pudo sostener en forma natural la fonación de la vocal. La duración del sonido varía entre 1 a 5 segundos. La base de datos original contiene diferentes fonaciones, en particular, en esta tesis, se utilizaron las grabaciones de la vocal /a/ sostenida. Se trabajó sobre 135 muestras de audio,

pertenecientes a las grabaciones seleccionadas habiendo de una a tres muestras por persona. De esta forma, se pudieron generar 135 espectrogramas en escala de grises, de los cuales 58 corresponden a enfermos y 77 a personas sanas. Para generar los espectrogramas, se utilizó Matlab R2018b. Se aplicó a las señales de audio de la base de datos la *Short-Time Fourier Transform (STFT)*, utilizando ventanas de 40 ms, con el solapamiento por defecto del 50 % y la escala de grises para representar la amplitud.

3.4. Aumentación por color

Para obtener un buen entrenamiento de las redes neuronales convolucionales es necesario contar con un gran volumen de datos. Dado qué el conjunto de sonidos seleccionados de la base de datos de audios de la vocal /a/, está formado por 135 muestras, se utilizó una primera estrategia para aumentar el número de espectrogramas La misma consiste en generar varios espectrogramas, a partir de la misma señal de audio, variando la paleta de colores utilizada.

Como ya se mencionó, un espectrograma permite representar a lo largo del tiempo las variaciones de frecuencia y amplitud de una señal de sonido. Es una representación en tres dimensiones: tiempo, frecuencia y amplitud. Comúnmente el espectrograma se representa a través de un gráfico en dos dimensiones: tiempo (eje horizontal) y frecuencia (eje vertical), donde la tercera dimensión (amplitud) es representada mediante el uso de una escala de colores.

Diferentes paletas de colores pueden resaltar diferentes aspectos del espectro de frecuencias y pueden dar lugar a diferentes percepciones y conclusiones. Por ejemplo, una paleta de colores que va del azul oscuro al blanco brillante puede resaltar las frecuencias de baja intensidad, mientras que una paleta que va del verde al rojo puede enfatizar las frecuencias de alta intensidad.

Además, algunas paletas de colores son más adecuadas para resaltar patrones y estructuras específicas en el espectrograma, como por ejemplo, una paleta de colores que se basa en el contraste de colores complementarios, puede ser

más útil para visualizar y analizar la estructura de los armónicos de una señal de audio. Por lo tanto, la elección de la paleta de colores puede influir en la interpretación de los datos y es importante tener en cuenta el propósito y la naturaleza de los datos que se están visualizando para elegir la paleta de colores más adecuada para cada caso.

Esto permite aumentar el número de muestras para el conjunto de entrenamiento de las redes, mejorar la precisión de clasificación y disminuir la variabilidad de los resultados.

Se crearon espectrogramas en escala de grises y con las paletas de color disponibles en la herramienta colormap de Matlab. Se consideraron 13 paletas de colores: “autumn”, “bone”, “cool”, “copper”, “gray”, “hot”,“hsv”, “jet”, “parula”, “pink”, “spring”, “summer” y “winter”. Las paletas “copper” y “bone” resaltan las formas de crestas y valles, mientras que “jet” o “hsv” da una indicación de la inclinación de las pendientes. Se excluyeron las paletas “colorcube”, “flag”, “lines”, “prism” y “white”, pues generan imágenes pixeladas y ruidosas, como se puede observar en la Figura 24.

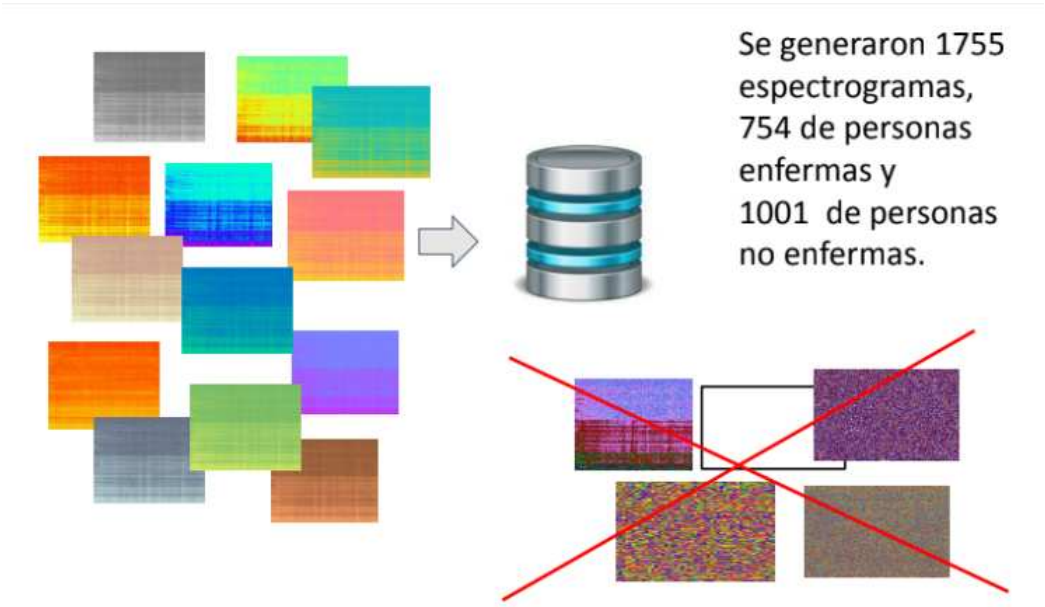


Figura 24: Elección de paletas de color

Considerando la estrategia de aumentación de datos, se generó una nueva base de datos que consta 1755 espectrogramas, 754 de personas enfermas y 1001 de personas sanas.

3.4.1. Clasificación entre enfermos y sanos Parkinson utilizando espectrogramas en color

Para mostrar los beneficios de la estrategia de aumentación de datos propuesta, se realizaron experimentos con diversos modelos de CNN que son representativos de las arquitecturas existentes: AlexNet [3], VGG 16 [115], ResNet 50 [6], Inception v3 [133] y Squeezenet [7]. En estos modelos se trabajó en modo de transferencia de aprendizaje, la misma permite con unas 500 a 1000 muestras por clase, entrenar modelos precisos.

Sobre todas las redes anteriores, se realizaron experimentos con dos conjuntos de datos: el primero con espectrogramas en escala de grises, y el segundo con el conjunto de datos aumentado considerando los espectrogramas en color.

Para obtener una medida de performance objetiva, se consideró un esquema de validación cruzada. Se realizaron 10 repeticiones del esquema de validación cruzada y se informó el promedio de las repeticiones.

Los códigos se desarrollaron con MatLab R2018b y Deep Learning Toolbox.

Se realizaron experimentos con diversos hiper-parámetros para las redes, de los cuales fueron seleccionados los siguientes:

- Conjunto Original: épocas 25, mini-batch 32, y razón de aprendizaje 0.0001.
- Conjunto Aumentado: épocas 35, mini-batch 32, y razón de aprendizaje 0.0001.

Los resultados del primer experimento de espectrogramas en escala de grises se muestran en la Tabla 3.1, donde las columnas se corresponden con el tipo de red neuronal utilizada (ALX: AlexNet, VGG: VGG16, IV3: Inception V3, RN5: ResNet50, SQZ: SqueezeNet), y las primeras 10 filas se corresponden con

una corrida del programa, donde se informa el promedio de las 10 repeticiones de la validación cruzada. Luego, se muestran los datos del ”Min”, el mínimo valor obtenido de las 10 corridas, ”Max” el máximo valor obtenido, ”Prom” el promedio de las 10 corridas, y ”Rango” es la diferencia entre el mejor y el peor valor obtenido.

Tabla 3.1: Resultados del Experimento realizado con espectrogramas en escala de grises, de los audios originales. Se presenta el % de accuracy obtenido en cada caso.

Corrida	ALX	VGG	IV3	RN5	SQZ
1	88,46 %	88,46 %	38,46 %	84,62 %	57,69 %
2	84,62 %	73,08 %	46,15 %	76,92 %	57,69 %
3	76,92 %	88,46 %	46,15 %	73,08 %	46,15 %
4	84,62 %	69,23 %	61,54 %	61,54 %	57,69 %
5	84,62 %	96,15 %	61,54 %	61,54 %	61,54 %
6	69,23 %	73,08 %	57,69 %	84,62 %	61,54 %
7	65,38 %	61,54 %	65,38 %	76,92 %	61,54 %
8	73,08 %	76,92 %	61,54 %	76,92 %	57,69 %
9	69,23 %	73,08 %	50,00 %	69,23 %	61,54 %
10	80,77 %	69,23 %	57,69 %	88,46 %	57,69 %
Min	65,38 %	61,54 %	38,46 %	61,54 %	46,15 %
Max	88,46 %	96,15 %	65,38 %	88,46 %	61,54 %
Prom	77,69 %	76,92 %	54,62 %	75,38 %	58,08 %
Rango	23,08 %	34,62 %	26,92 %	26,92 %	15,38 %

Los experimentos realizados con el conjunto de datos de espectrogramas en escala de grises arrojaron resultados poco confiables debido a la gran variabilidad de los mismos.

En el gráfico de la izquierda de la Figura 25, se visualizan, a modo de ejemplo, la performance obtenida en cada una de las 10 corridas realizadas sobre la red de arquitectura VGG16. Se puede observar que existe una gran variabilidad en los resultados. Lo mismo sucedió para las otras arquitecturas utilizadas, como se presenta en el diagrama de cajas de la derecha, donde se observa que para todas las redes, la dispersión fue grande, salvo en SqueezeNet, pero su tasa promedio de acierto estuvo en el 58,08 %, y el máximo solo llegó al 61,54 %. La red Inception V3 tuvo muy malos resultados para este conjunto de datos: su dispersión fue alta con un 26.92 %, pero su máximo solo llegó al 65.38 %, y su mínimo estuvo en un 38.46 %, el peor valor obtenido para todas las redes.

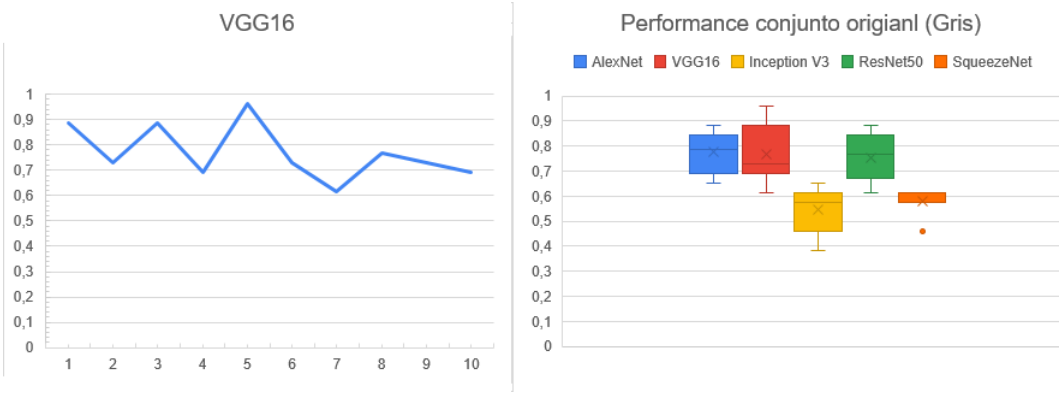


Figura 25: Resultados de los espectrogramas en escala de grises

La Tabla 3.2 presenta los resultados obtenidos para cada red. Se informa el máximo, el mínimo, la distancia entre el máximo y el mínimo (Rango), y el promedio de las repeticiones de la clasificación del conjunto de test. Las columnas "Orig" corresponden a los resultados con los espectrogramas en escala de grises originales, y las columnas "Aum" corresponden a los resultados del conjunto de datos aumentado. También se observa, que para todos los modelos de redes, la estrategia de aumentación de datos permitió mejorar tanto la razón de acierto como disminuir la dispersión de los resultados. La red VGG16 obtuvo los mejores indicadores, con un promedio en la tasa de acierto de 95.98 %, un máximo de 98.01 %, y un mínimo de 92.88 %.

Tabla 3.2: Comparación del % de acierto obtenido en en la clasificación del conjunto de test, entre datos originales y aumentados, con diversos modelos de CNN.

	Máximo % acierto		**Mínimo % acierto**		**Rango % acierto**		**Promedio % acierto**	
	Orig	**Aum**	**Orig**	**Aum**	**Orig**	**Aum**	**Orig**	**Aum**
AlexNet	88.46 %	91.74 %	65.38 %	81.20 %	23.08 %	10.54 %	77.69 %	87.64 %
VGG16	96.15 %	98.01 %	61.54 %	92.88 %	34.62 %	5.13 %	76.92 %	95.98 %
Inception V3	65.38 %	86.89 %	38.46 %	78.92 %	26.92 %	7.98 %	54.62 %	83.13 %
ResNet50	88.46 %	90.03 %	61.54 %	86.89 %	26.92 %	3.13 %	75.38 %	88.09 %
SqueezeNet	61.54 %	84.05 %	46.15 %	73.79 %	15.38 %	10.26 %	58.08 %	80.09 %

En la Figura 26 se puede ver la comparación de los dos conjuntos de datos utilizados. El gráfico de la izquierda, corresponde al conjunto de espectrogramas en escala de grises detallado anteriormente, donde se puede observar que para todas las redes, la dispersión fue grande, El gráfico de la derecha muestra los resultados de los experimentos sobre el conjunto aumentado con los 13

colores de espectrogramas. En este caso se observa que para todas las redes la dispersión disminuyó. Y el promedio de aciertos mejoró obteniéndose para todos los casos un promedio de aciertos superior al 80 %.

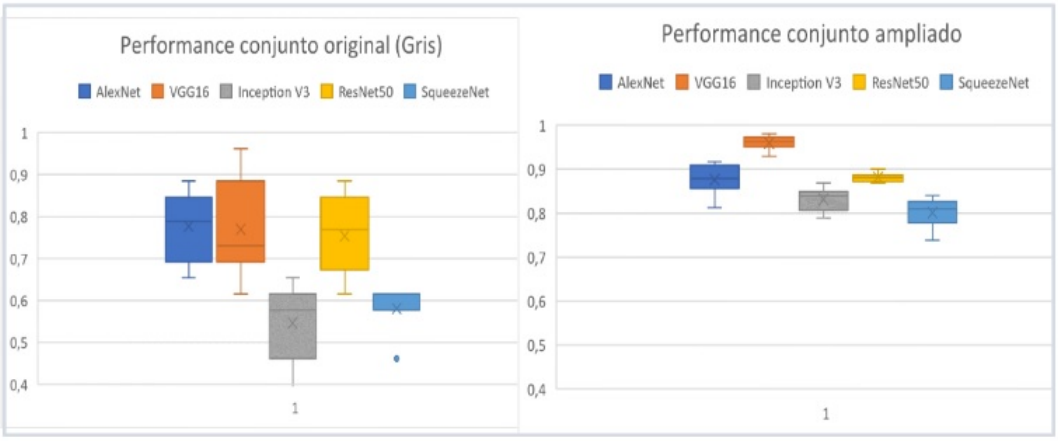


Figura 26: Diferencia de variabilidad entre los resultados utilizando espectrogramas en escala de grises y a color

3.5. Aumentación por segmentación del audio

En este caso, se decidió utilizar una estrategia para aumentar la cantidad de muestras de audio a partir de las cuales, generar los espectrogramas tanto en escala de grises como en las 13 paletas de colores seleccionadas.

Esta segunda metodología para aumentar la cantidad de muestras consistió en generar fragmentos de los sonidos originales. Los 135 sonidos originales tenían una duración variable de 3 a 5 segundos. Se realizaron cortes de 1 segundo con 50 % de solapamiento de los sonidos originales. En la Figura 27 se puede observar este procedimiento para un sonido de 2 segundos, del cual se obtienen 3 fragmentos de 1 segundo cada uno. Con esta estrategia de aumentación de datos se obtuvieron 1168 muestras de sonido de 1 segundo cada una. Para cada uno de estos fragmentos se generaron los espectrogramas en escala de grises y en las 13 paletas de color de MatLab seleccionadas. De esta forma, se obtuvieron 1168 espectrogramas en escala de grises y 15.184 espectrogramas en color.

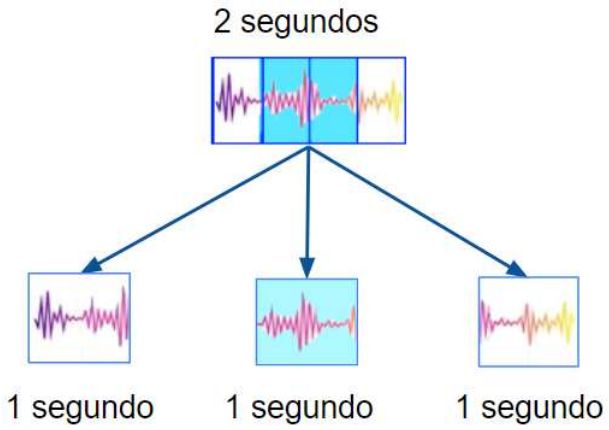


Figura 27: Procedimiento de corte de un sonido de 2 segundos con un solapamiento del 50 %.

3.6. Experimentos

3.6.1. El esquema general del proceso de trabajo

El esquema general del proceso de trabajo planteado en esta tesis se muestra en la Figura 28. Se inició con los sonidos de la base de datos de audios de voz, los cuales se transformaron en espectrogramas utilizando la STFT. Estos espectrogramas se utilizaron para crear una nueva base de datos a la que se le aplicaron técnicas de Deep Learning para clasificar entre enfermos y no enfermos de Parkinson. La información obtenida a partir de este análisis proporcionará un valioso soporte al médico para realizar el diagnóstico y seguimiento de la enfermedad.

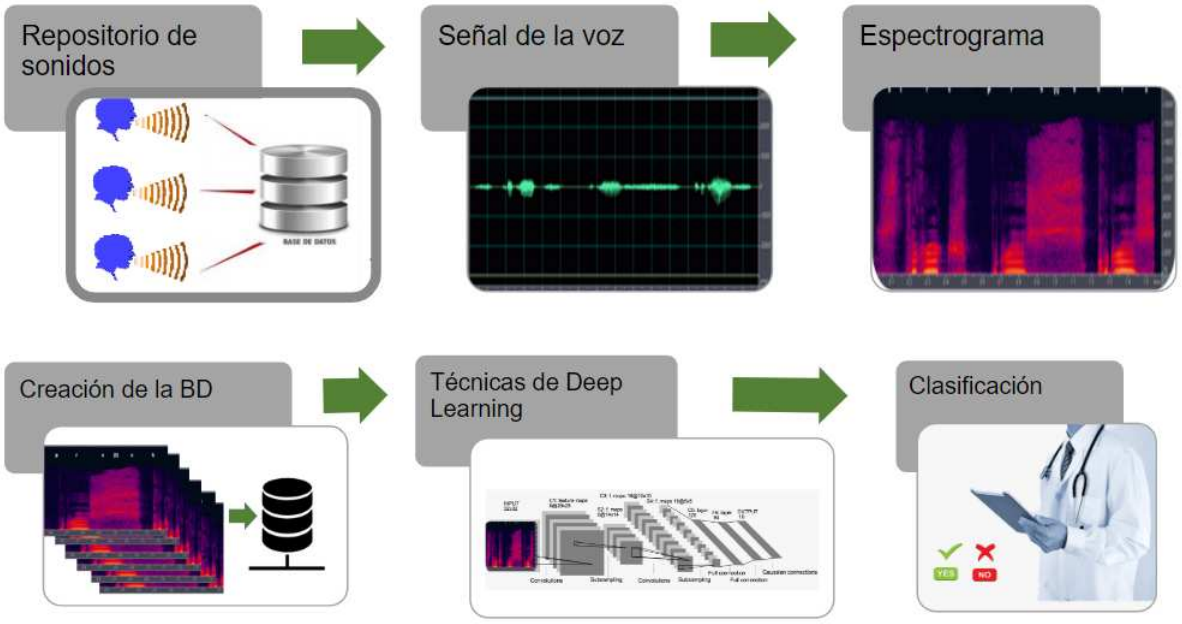


Figura 28: Proceso General de Trabajo utilizado en esta Tesis

Se realizaron experimentos considerando espectrogramas con sonidos originales y con fragmentos de sonidos. Los sonidos originales son aquellos cuya duración corresponde al tiempo que la persona pudo sostener en forma natural la fonación de la vocal. Los fragmentos son cortes de 1 segundo con 50 % de solapamiento de los sonidos originales. Los espectrogramas en escala de grises corresponden con una única matriz, resultante de aplicar la STFT a la señal de sonido, donde los valores de cada celda (píxel de la imagen) varía de entre 0 (negro) y 255 (blanco). Los espectrogramas en color se generaron utilizando las 13 paletas de colores seleccionadas de MatLab (incluyendo la paleta en escala de grises). Toda imagen en color expresada en el sistema RGB (red, green and blue), tiene 3 canales: rojo, verde y azul, cada uno de los cuales se corresponde con una matriz donde cada celda tiene valores entre 0 y 255. Las imágenes en escala de grises de un solo canal deben estar previamente procesadas como imágenes pseudo-color de 3 canales (se repite en los 3 canales la misma matriz) antes de que la red pueda analizarlas. Al trabajar en los experimentos con la técnica de transferencia de aprendizaje, un modelo previamente entrenado con imágenes en color introduce artefactos, e ineficiencias, en las imágenes de un solo canal [151].

Considerando lo anterior, se realizaron 4 tipos de experimentos:

- Experimento 1 (EXP1): Con espectrogramas en escala de grises de sonidos originales.
- Experimento 2 (EXP2): Con espectrogramas en color de sonidos originales.
- Experimento 3 (EXP3): Con espectrogramas en color de fragmentos de sonido.
- Experimento 4 (EXP4): Con espectrogramas en color de sonidos originales y fragmentos.

La Figura 29 presenta un esquema gráfico de los conjuntos de datos utilizados en cada experimento.

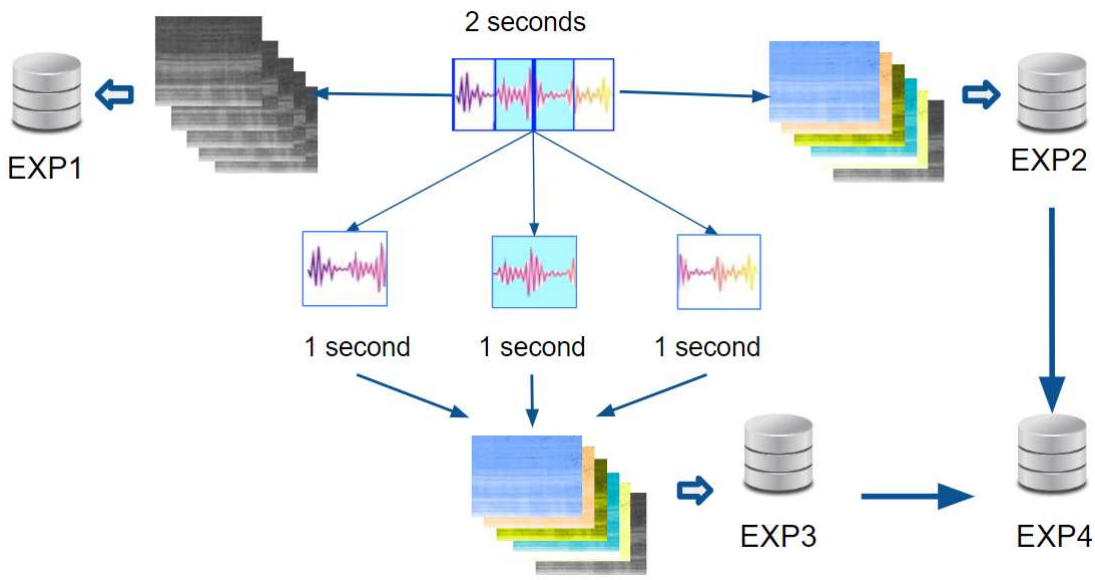


Figura 29: Muestras de audio que forman las bases de datos de cada uno de los 4 experimentos considerados.

3.6.2. Proceso general de los experimentos

En esta tesis se trabaja con dos modelos de aprendizaje profundo, diseñados para clasificar EP a partir de espectrogramas obtenidos de las grabaciones de la fonación de la vocal /a/ sostenida. Uno es el modelo de aprendizaje con CNN, y el otro es el modelo de aprendizaje basado en ELM. En la Figura 30 se presentan estos dos modelos de aprendizaje.

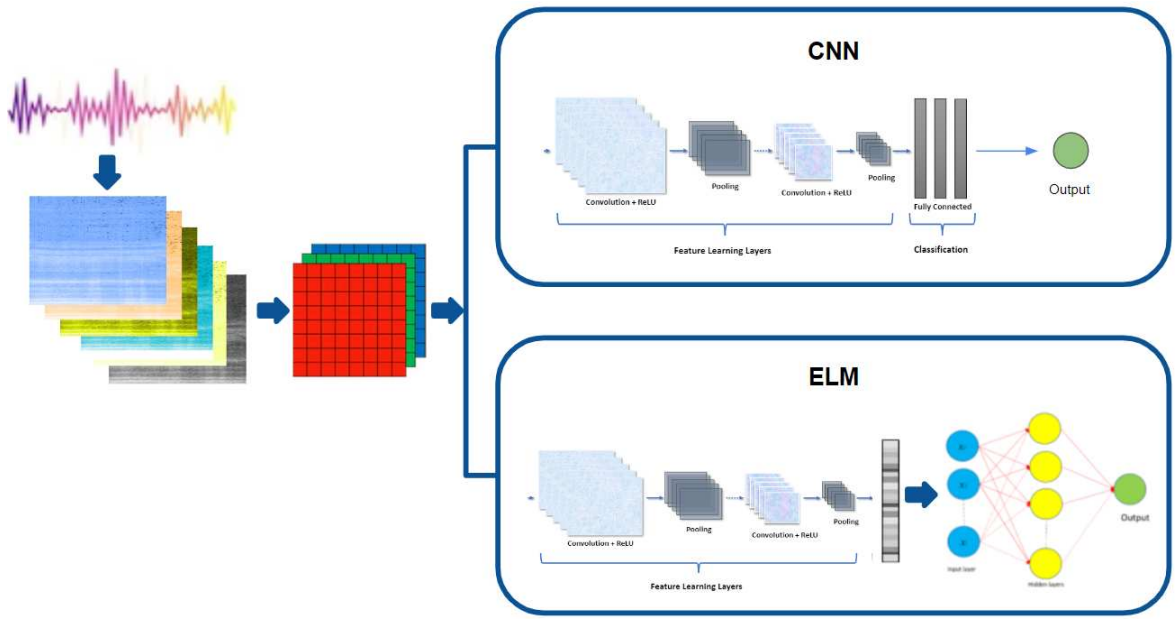


Figura 30: Modelos de aprendizaje profundo para la clasificación de EP.

Como se puede observar, el proceso parte de las señales de sonido convertidas en espectrogramas. La entrada al modelo de CNN es la representación matricial de los espectrogramas (3 canales del modelo RGB). La red extrae en forma automática las características en el período de entrenamiento, y utiliza las últimas capas para realizar la clasificación entre enfermos y no enfermos de Parkinson. La entrada al modelo ELM son las características extraídas al aplicar una CNN a los espectrogramas. Es decir que, de esta CNN, no se realiza la etapa de clasificación, sino que se aplica una ELM al vector de características obtenido de ella.

Las arquitecturas de CNN utilizadas en todos los experimentos son AlexNet, Squeezenet, Inception V3, ResNet50 y VGG16.

3.6.3. Esquema de validación cruzada

La técnica de validación cruzada se implementó para garantizar la fiabilidad de los resultados de todos los experimentos. La misma se llevó a cabo mediante la generación de 10 lotes (lote 0 a lote 9), cada uno compuesto por tres conjuntos disjuntos de personas: 70 % para el conjunto de entrenamiento, 10 % para conjunto de validación y 20 % para conjunto de test. El esquema utilizado se muestra en la Figura 31.

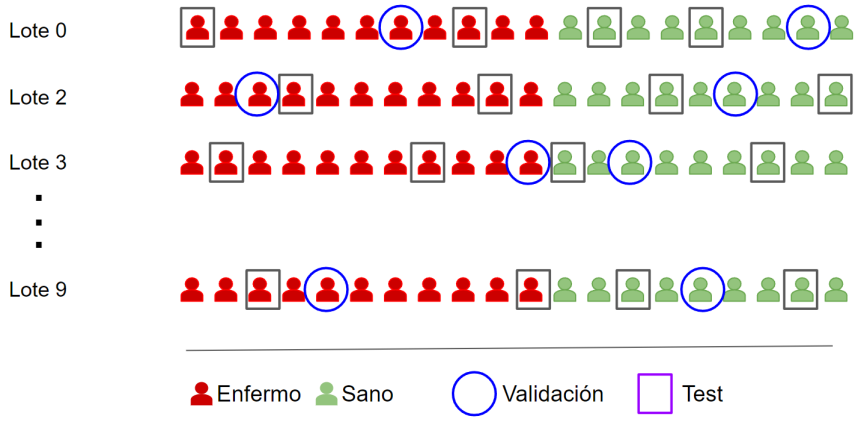


Figura 31: Validación cruzada

De cada persona se incluyeron todas las muestras, según el experimento: sonidos originales, fragmentos o ambos. Estos lotes permitieron llevar a cabo experimentos tanto en redes convolucionales (CNN) como en máquinas de aprendizaje extremo (ELM), utilizando las mismas muestras de datos para ambos. De esta forma, se evitó que los resultados dependieran de la selección de muestras particulares, asegurando que las conclusiones fueran robustas y generalizables.

3.6.4. Esquema General de proceso de clasificación con CNN

El proceso general de clasificación de enfermos de Parkinson utilizando CNNs es el mismo para todas las arquitecturas consideradas, como ilustra la Figura 32.

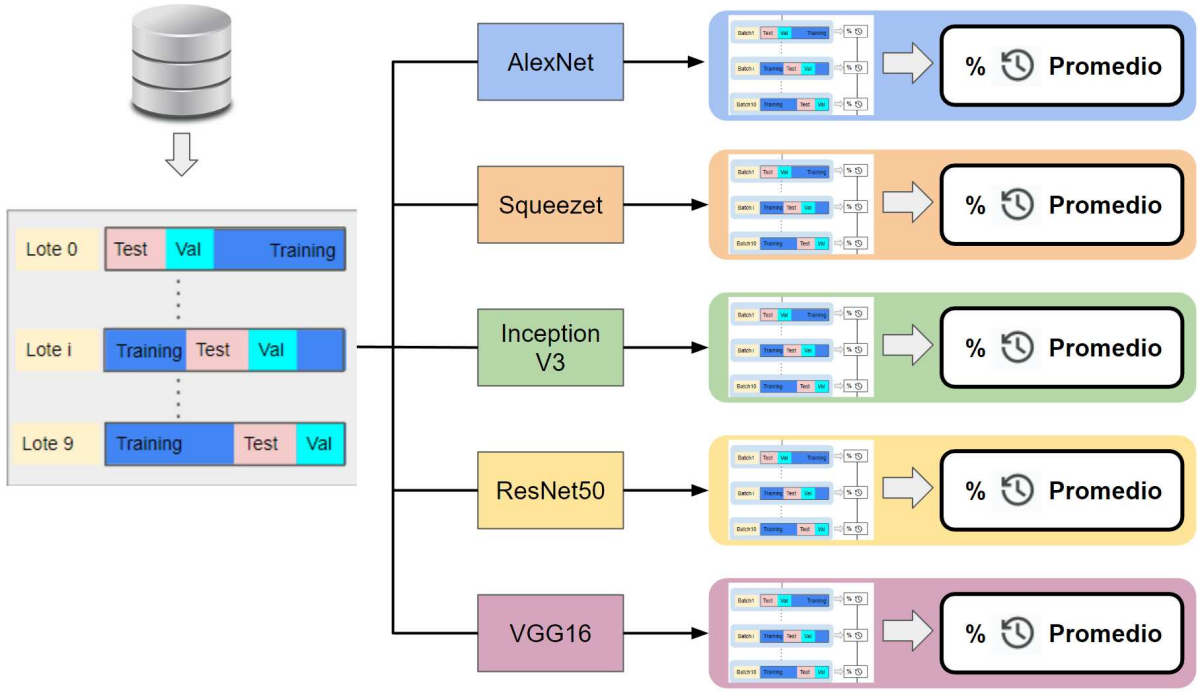


Figura 32: Esquema general del proceso de clasificación con CNN

Debido al número de muestras disponibles, se utilizó la técnica de transferencia de aprendizaje para las distintas arquitecturas de CNN, consideradas en este trabajo. Se utilizó una CNN previamente entrenada en un conjunto de datos grande y complejo (ImageNet), y se ajustó para el conjunto de datos más pequeño y específico del lote de datos. Para ello, se eliminaron las capas finales de la red y se añadieron nuevas capas.

Para cada lote se entrenó el modelo con el conjunto de entrenamiento y validación que contenían respectivamente el 70 % y el 10 % de los datos. De esta forma, se obtuvo el modelo entrenado. Este proceso se repitió 10 veces, lo que permitió obtener 10 modelos entrenados. De estos 10 modelos, se seleccionó el mejor para este lote. Finalmente, se utilizó este mejor modelo para evaluar el conjunto de datos de prueba, obteniéndose así la precisión (*accuracy*), y el tiempo de entrenamiento del lote.

Luego de repetir el proceso para cada uno de los 10 lotes, se obtuvo la precisión de la red como el promedio de las precisiones de cada lote. Así mismo se registró el tiempo de entrenamiento, de validación, y de test de cada lote, para calcular e informar el promedio de tiempo en cada arquitectura.

La Figura 33 ilustra el proceso descripto llevado a cabo para cada una de las CNN seleccionadas.

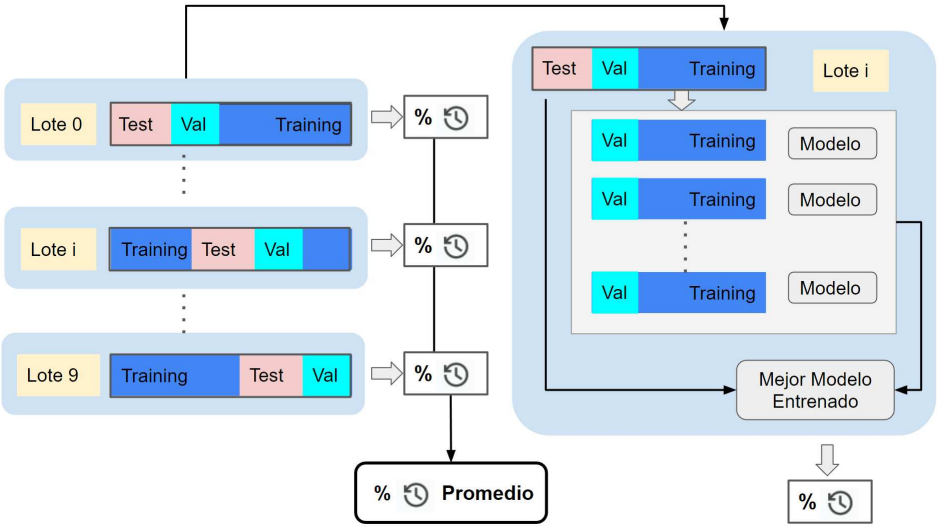


Figura 33: Proceso General para una arquitectura CNN particular

3.6.5. Esquema General del Clasificador ELM

Para aplicar el clasificador ELM, en primer lugar, se crearon los vectores de características obtenidos para cada una de las CNN consideradas, ver Figura 34.

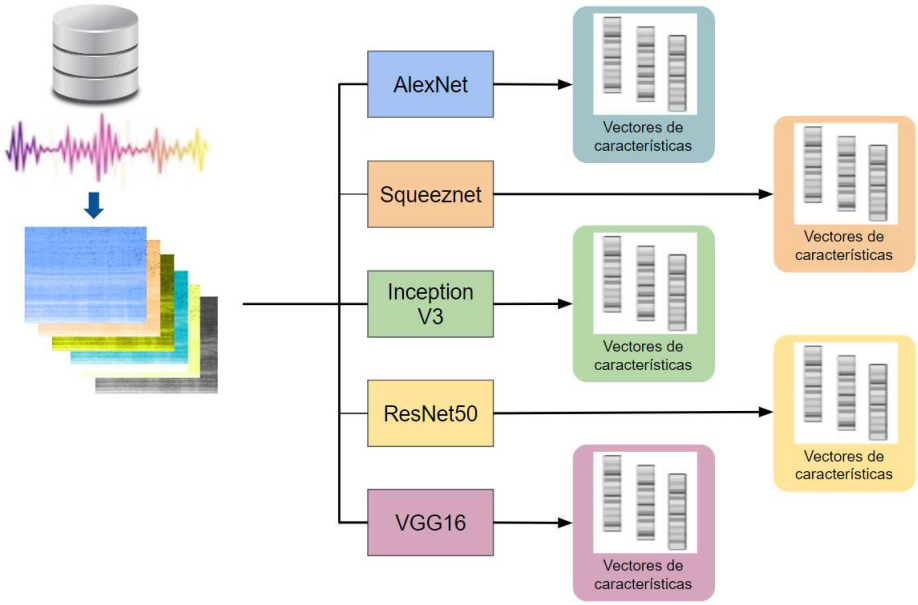


Figura 34: Proceso de obtención de los vectores de características por arquitectura CNN particular

Estos vectores de características se obtuvieron al pasar las imágenes de todos los espectrogramas de la base de datos por la CNN y quedarse con una capa intermedia de la red. Cada vector de características codifica información relevante sobre las características visuales de la imagen de entrada.

Para cada lote, se seleccionaron los vectores de características correspondientes a las muestras de ese lote en particular. De esta forma, se mantuvieron las características constantes para todos los experimentos realizados, lo que permitió la comparación entre ellos. Si las características no se mantuvieran constantes, se produciría una variación en cada entrenamiento y no se podría realizar una comparación significativa.

En resumen, la idea fue utilizar las características obtenidas a partir de una capa intermedia de la CNN para alimentar al clasificador ELM, con el fin de realizar la clasificación de enfermos y no enfermos de Parkinson.

Al igual que en el caso anterior, se registró el accuracy y tiempos por lote para informar los promedios por cada arquitectura.

En la Figura 35 se puede observar el proceso antes descripto.

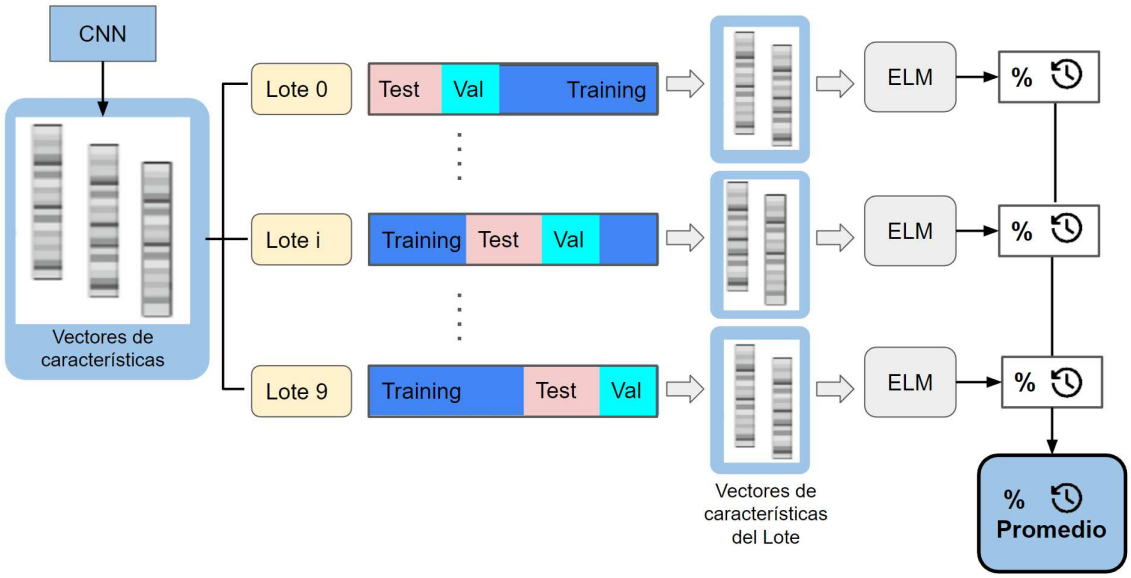


Figura 35: Proceso General de clasificación de ELM utilizando los vectores de características obtenidos de una arquitectura CNN particular

3.6.6. Computo de los hiper-parámetros

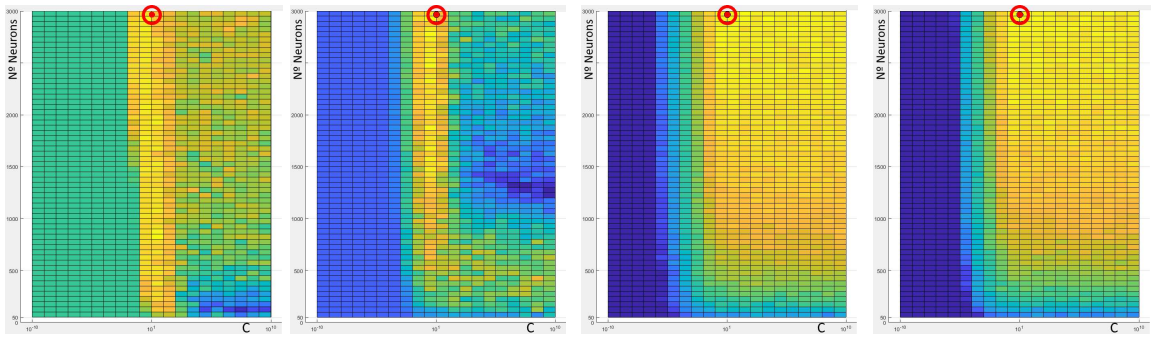
Se conoce como hiperparámetros, a aquellos valores que no se aprenden durante el entrenamiento, sino que deben ser establecidos antes de comenzar el proceso de entrenamiento de la red. Estos influyen en cómo se ajustan los parámetros del modelo y, por lo tanto, afectan directamente su rendimiento y capacidad de generalización.

Para la adecuada búsqueda de los hiper-parámetros, se realizaron experimentos con diversas combinaciones de épocas, mini-batch y tasa de aprendizaje, para los distintos experimentos y CNN consideradas, de los cuales fueron seleccionados los siguientes:

- EXP1: épocas 25, mini-batch 32 y razón de aprendizaje 0.0001.
- EXP2, EXP3, EXP4: épocas 35, mini-batch 32 y razón de aprendizaje 0.0001.

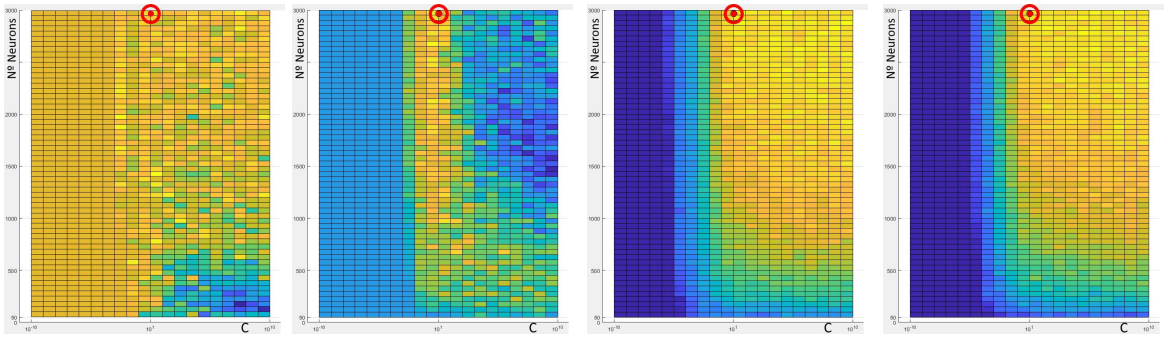
Para las ELM, se realizó una serie de pruebas para establecer los hiper-parámetros óptimos. Para cada una de las arquitecturas consideradas en los distintos experimentos (EXP1, EXP2, EXP3 y EXP4), se realizaron distintas

iteraciones combinando la variación de la cantidad de neuronas de la capa oculta de 1 a 3000 y el parámetro de regularización entre 10^{-10} y 10^{10}. Para las arquitecturas AlexNet, Incepction V3, ResNet 50, SqueezNet, VGG 16, las superficies de hiperparámetros se muestran en las Figuras 36, 37, 38, 39 y 40, respectivamente. En cada imagen, el eje X corresponde con el parámetro de regularización y el eje Y con el número de neuronas, el acuracy obtenido está representado por el color de la imagen, siendo el amarillo los valores más altos y los azules los más bajos. De la comparación de estás imágenes se puede observar que los valores de los hiperparámetros óptimos encontrados para el EXP1, 3000 neuronas y el parámetro de regularización 10^1, son los mismos para los otros experimentos. El set de hiperparámetros óptimos se muestra en las imágenes con un punto rojo.



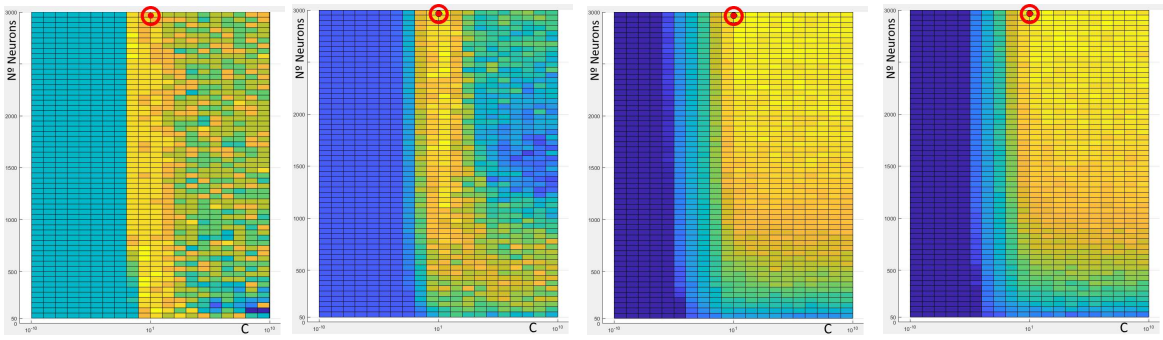
(a) EXP1 (b) EXP2 (c) EXP3 (d) EXP4

Figura 36: Búsqueda de los hiperparámetros para AlexNet: número de neuronas y el parámetro de regularización *"c"*



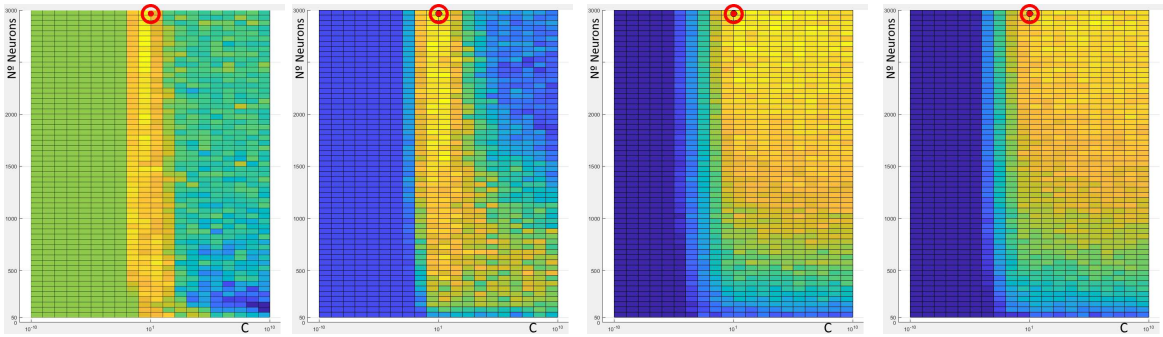
(a) EXP1 (b) EXP2 (c) EXP3 (d) EXP4

Figura 37: Búsqueda de los hiperparámetros para Incepction V3: número de neuronas y el parámetro de regularización *"c"*



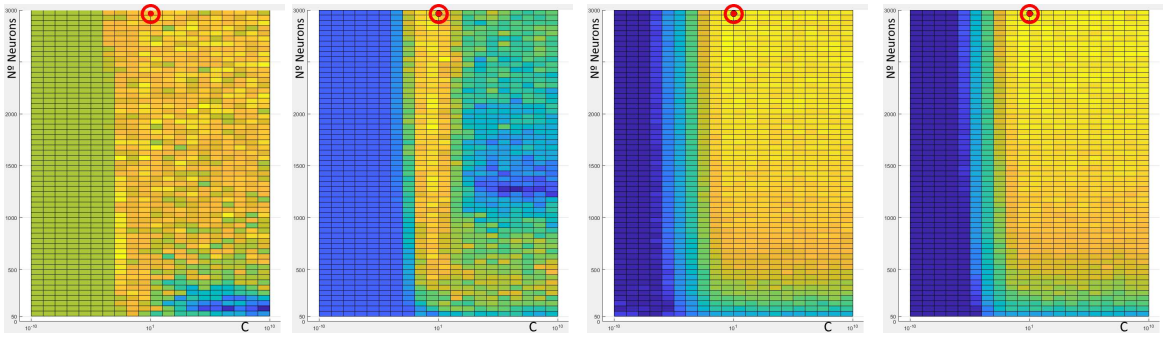
(a) EXP1 (b) EXP2 (c) EXP3 (d) EXP4

Figura 38: Búsqueda de los hiperparámetros para ResNet 50: número de neuronas y el parámetro de regularización *"c"*



(a) EXP1 (b) EXP2 (c) EXP3 (d) EXP4

Figura 39: Búsqueda de los hiperparámetros para SqueezNet: número de neuronas y el parámetro de regularización *"c"*



(a) EXP1 (b) EXP2 (c) EXP3 (d) EXP4

Figura 40: Búsqueda de los hiperparámetros para VGG 16: número de neuronas y el parámetro de regularización *"c"*

En todos los casos, se ha utilizado como función de activación la función sigmoidea. La elección de la función sigmoidea se debe a que la misma permite que una red neuronal modele relaciones no lineales complejas entre entradas y salidas. Esta propiedad es fundamental para resolver problemas no lineales como la clasificación, el modelado del lenguaje, y el reconocimiento de voz. Además, su salida está entre 0 y 1, lo que lo hace útil en aplicaciones de clasificación binaria, que es nuestro caso ya que queremos determinar si un paciente está enfermo o no [152]. La función de activación utilizada es la siguiente:

$$g(w_i x_j + b_i) = \frac{1}{1 + e^{-(w_i x_j + b_i)}} \tag{8}$$

donde w_i y b_i son respectivamente los pesos y biases de la capa oculta, y x_j los datos de entrada a la red.

Página intencionalmente en blanco

Capítulo 4

Resultados y análisis

4.1. Resultados

En esta sección se presentan diferentes tablas que permiten visualizar los resultados obtenidos para todas las redes neuronales utilizadas en cada uno de los experimentos realizados. Las tablas presentadas son: de accuracy, de tiempo de entrenamiento y test, y de comparación de métricas de clasificación.

En las tablas de accuracy, cada fila corresponde al promedio de performance para cada uno de los 10 lotes considerados. Al final de la tabla, se indican los valores máximo, mínimo, promedio y rango de variación entre el máximo y el mínimo obtenido. En las tablas de tiempo, cada fila se corresponde con el tiempo promedio de entrenamiento y test de cada lote considerado. La última fila muestra el tiempo promedio de entrenamiento y test de todos los lotes. En la tabla de métricas de clasificación, se pueden visualizar los promedios de recall (TPR), especificidad (TNR), y accuracy (ACC), para los 10 lotes considerados.

El accuracy que se muestra indica el porcentaje de aciertos por persona. Vale aclarar que todos los lotes de test están formados por las muestras de 23 personas. Para calcular el accuracy de una persona, se consideró como acierto si más del 50 % de sus muestras, resultaron bien clasificadas. Al ser tan pocas personas, si una de ellas resulta mal clasificada, el porcentaje de acierto del lote desciende del 100 % (0 error), hasta 95.65 % (1 error).

En las tablas de tiempo se puede observar que el tiempo aumenta significativamente cuando se utilizan las redes Inception V3, ResNet50 y VGG16, en comparación con modelos menos complejos como AlexNet y SqueezeNet. La complejidad de la red está dada por la cantidad de capas, de filtros y parámetros. Como se puede observar en la Tabla 2.1, las redes Inception V3, ResNet50 y VGG16 tienen mayor cantidad de capas, filtros y parámetros que AlexNet y SqueezNet. Esto implica una mayor cantidad de cálculos, lo que se ve reflejado en un aumento significativo del tiempo de entrenamiento. Este incremento es mucho más notorio en los tiempos de entrenamiento, debido a la cantidad de muestras procesadas.

En la Tabla 4.1 se presentan las abreviaturas de las distintas arquitecturas utilizadas en la tablas de resultados de los experimentos realizados.

Tabla 4.1: Abreviaturas para las CNNs utilizadas

Arquitectura considerada	**Abreviatura utilizada**
AlexNet	ALX
Squeezenet	SQZ
Inception V3	IV3
ResNet-50	RN5
VGG-16	VGG

4.1.1. Experimento 1: Espectrogramas en escala de grises de sonidos originales

Este experimento está formado por un total de 135 muestras para las etapas de entrenamiento, validación y test, correspondientes a 23 personas. Es el experimento realizado con menor cantidad de muestras.

La Tabla 4.2 muestra que las ELM tuvieron un rendimiento similar a las CNN, en cuanto al accuracy, llegando a un máximo de 91.30 %. La Tabla 4.3 muestra que, el tiempo de entrenamiento para las ELM estuvo en un rango de 7 a 7.9 seg, mientras que las CNN entre 5.6 y 38.4 seg. Los tiempos de test fueron similiares en ambas redes, con valores no relevantes. Aunque el número de muestras es reducido, se comienza a evidenciar que las ELM tienen un menor

tiempo de entrenamiento.

En la Tabla 4.4 se puede observar que los valores de TPR, para las CNN, son muy bajos, mejorando para las ELM en todas las arquitecturas consideradas. En general se puede observar que clasifica mejor a los sujetos sanos que los enfermos de Parkinson.

Tabla 4.2: Experimento 1: Accuracy (135 muestras)

	ALX		**SQZ**		**IV3**		**RN5**		**VGG**	
	CNN	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**
Lote0	73.91%	78.26%	69.57%	69.57%	60.87%	86.96%	78.26%	86.96%	73.91%	78.26%
Lote1	65.22%	73.91%	60.87%	69.57%	43.48%	65.22%	56.52%	73.91%	69.57%	65.22%
Lote2	69.57%	69.57%	60.87%	86.96%	47.83%	60.87%	78.26%	65.22%	86.96%	69.57%
Lote3	73.91%	86.96%	60.87%	65.22%	56.52%	60.87%	73.91%	78.26%	82.61%	82.61%
Lote4	78.26%	73.91%	39.13%	65.22%	56.52%	69.57%	73.91%	82.61%	78.26%	73.91%
Lote5	82.61%	86.96%	39.13%	65.22%	56.52%	60.87%	82.61%	86.96%	82.61%	78.26%
Lote6	82.61%	82.61%	39.13%	73.91%	65.22%	65.22%	78.26%	73.91%	78.26%	82.61%
Lote7	86.96%	91.30%	60.87%	73.91%	73.91%	65.22%	73.91%	82.61%	82.61%	86.96%
Lote8	65.22%	82.61%	39.13%	65.22%	65.22%	78.26%	73.91%	86.96%	69.57%	65.22%
Lote9	82.61%	73.91%	39.13%	86.96%	60.87%	69.57%	91.30%	78.26%	82.61%	82.61%
MAX	86.96%	**91.30%**	69.57%	86.96%	73.91%	86.96%	**91.30%**	86.96%	86.96%	86.96%
MIN	65.22%	69.57%	39.13%	65.22%	43.48%	60.87%	56.52%	65.22%	69.57%	65.22%
PROM	76.09%	80.00%	50.87%	72.17%	58.70%	68.26%	76.09%	79.57%	78.70%	76.52%
RANGO	21.74%	21.74%	30.43%	21.74%	30.43%	26.09%	34.78%	21.74%	17.39%	21.74%

Tabla 4.3: Experimento 1: Tiempos de entrenamiento y test (135 muestras)

Lote	**TIEMPO DE ENTRENAMIENTO**										**TIEMPO DE TEST**									
	ALX		**SQZ**		**IV3**		**RN5**		**VGG**		**ALX**		**SQZ**		**IV3**		**RN5**		**VGG**	
	CNN	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**
L0	5,3	7,5	6,7	6,8	38,2	7,2	24,2	6,8	23,4	7,6	0,2	0,2	0,2	0,1	0,4	0,1	0,3	0,1	0,4	0,3
L1	5,5	7,7	6,7	7,2	38,3	6,9	24,8	6,8	24,2	8,3	0,2	0,3	0,2	0,1	0,3	0,1	0,2	0,1	0,3	0,3
L2	5,5	7,5	6,7	7,0	39,0	7,0	24,9	6,9	24,5	7,5	0,2	0,3	0,2	0,1	0,3	0,1	0,2	0,1	0,3	0,2
L3	5,7	7,6	6,7	7,1	38,1	7,1	25,0	7,4	24,3	7,8	0,2	0,3	0,2	0,1	0,4	0,1	0,3	0,1	0,4	0,4
L4	5,5	8,1	6,7	7,3	38,5	7,5	24,5	6,9	24,7	8,8	0,2	0,4	0,2	0,1	0,4	0,1	0,3	0,1	0,4	0,2
L5	5,9	7,9	6,8	7,1	39,6	6,8	24,8	6,8	24,5	7,4	0,2	0,3	0,2	0,1	0,3	0,1	0,3	0,1	0,4	0,2
L6	5,7	7,4	7,0	6,9	38,5	7,4	24,8	7,0	24,6	7,7	0,2	0,3	0,2	0,1	0,3	0,1	0,3	0,1	0,3	0,3
L7	5,8	7,9	6,9	6,8	37,9	7,2	25,6	7,0	24,4	7,5	0,2	0,3	0,2	0,1	0,3	0,1	0,2	0,1	0,3	0,4
L8	5,6	7,6	6,8	7,3	38,1	7,0	24,7	7,3	24,4	7,7	0,2	0,3	0,2	0,1	0,3	0,1	0,3	0,1	0,3	0,5
L9	5,7	7,6	6,7	7,0	38,3	8,6	24,8	6,9	24,2	9,0	0,2	0,3	0,2	0,1	0,3	0,1	0,3	0,1	0,4	0,3
PROM	5,6	7,7	6,8	7,0	38,4	7,3	24,8	7,0	24,3	7,9	0,2	0,3	0,2	0,1	0,3	0,1	0,3	0,1	0,4	0,3

Tabla 4.4: Experimento 1: Comparación entre TPR - TNR- ACC (135 muestras)

	ALX		**SQZ**		**IV3**		**RN5**		**VGG**	
	CNN	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**
TPR	57,78%	74,44%	54,44%	55,56%	47,78%	60,00%	52,22%	74,44%	63,33%	72,22%
TNR	87,86%	83,57%	48,57%	82,86%	65,71%	73,57%	91,43%	82,86%	88,57%	79,29%
ACC	76,09%	80,00%	50,87%	72,17%	58,70%	68,26%	76,09%	79,57%	78,70%	76,52%

4.1.2. Experimento 2: Espectrogramas en color sonidos originales

Con el fin de aumentar la cantidad de muestras, se realizó una aumentación de datos generando espectrogramas con distintas escalas de colores, obteniéndose de esta forma un total de 1754 muestras.

En la Tabla 4.5 de accuracy se puede observar que tanto el máximo obtenido con las CNN y ELM fue del 95.65 %, que es el máximo que se puede obtener con una única persona mal clasificada. En general, se puede observar que los valores de Acuraccy se mantienen similares entre los dos tipos de redes, y que los tiempos de entrenamiento son menores para las ELM.

Tabla 4.5: Experimento 2: Accuracy (1.754 muestras)

	ALX		**SQZ**		**IV3**		**RN5**		**VGG**	
	CNN	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**
Lote0	91.30 %	95.65 %	60.87 %	82.61 %	69.57 %	78.26 %	82.61 %	86.96 %	86.96 %	78.26 %
Lote1	82.61 %	86.96 %	69.57 %	73.91 %	65.22 %	65.22 %	73.91 %	82.61 %	65.22 %	73.91 %
Lote2	78.26 %	69.57 %	73.91 %	82.61 %	69.57 %	65.22 %	86.96 %	82.61 %	82.61 %	73.91 %
Lote3	78.26 %	69.57 %	65.22 %	60.87 %	73.91 %	69.57 %	73.91 %	69.57 %	69.57 %	65.22 %
Lote4	73.91 %	78.26 %	69.57 %	86.96 %	65.22 %	65.22 %	73.91 %	69.57 %	73.91 %	78.26 %
Lote5	86.96 %	86.96 %	60.87 %	86.96 %	56.52 %	91.30 %	91.30 %	86.96 %	91.30 %	86.96 %
Lote6	91.30 %	95.65 %	78.26 %	82.61 %	73.91 %	78.26 %	78.26 %	82.61 %	91.30 %	86.96 %
Lote7	95.65 %	91.30 %	60.87 %	82.61 %	73.91 %	73.91 %	91.30 %	95.65 %	69.57 %	91.30 %
Lote8	60.87 %	69.57 %	60.87 %	60.87 %	60.87 %	69.57 %	73.91 %	78.26 %	78.26 %	65.22 %
Lote9	95.65 %	82.61 %	82.61 %	91.30 %	69.57 %	65.22 %	86.96 %	82.61 %	91.30 %	91.30 %
MAX	95.65 %	**95.65 %**	82.61 %	91.30 %	73.91 %	91.30 %	**91.30 %**	95.65 %	91.30 %	91.30 %
MIN	60.87 %	69.57 %	60.87 %	60.87 %	56.52 %	65.22 %	73.91 %	69.57 %	65.22 %	65.22 %
PROM	83.48 %	82.61 %	68.26 %	79.13 %	67.83 %	72.17 %	81.30 %	81.74 %	80.00 %	79.13 %
RANGO	34.78 %	26.09 %	21.74 %	30.43 %	17.39 %	26.09 %	17.39 %	26.09 %	26.09 %	26.09 %

A partir del análisis de la Tabla 4.6, se observa que los tiempos de entrenamiento de las ELM varian entre 8.4 a 9.9 seg, mientras que para las CNN varían entre 50.7 a 569.8 seg. Los tiempos de test continúan siendo similares entre ambas redes.

En la Tabla 4.7 se puede observar que para todos los casos se obtiene una buena clasificación de los sanos, alcanzando un valor máximo de TNR del 97.89 %. En relación con los valores obtenidos en el experimento 1, se puede apreciar un aumento significativo de la tasa TNR tanto para las CNN y ELM, mientras que el TPR no tiene una mejora significativa para las CNN y empeora para las ELM.

Tabla 4.6: Experimento 2: Tiempos de entrenamiento y test (1.754 muestras)

Lote	TIEMPO DE ENTRENAMIENTO										TIEMPO DE TEST									
	ALX		SQZ		IV3		RN5		VGG		ALX		SQZ		IV3		RN5		VGG	
	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM
L0	51,3	9,8	62,4	8,5	576,8	8,2	359,0	8,7	396,0	9,6	0,6	0,7	0,6	0,2	2,3	0,2	1,8	0,2	2,3	0,5
L1	51,8	9,3	62,0	8,4	576,6	8,3	364,2	8,4	416,3	9,1	0,6	0,6	0,6	0,2	2,1	0,2	1,7	0,2	2,3	0,5
L2	50,7	10,1	62,2	8,5	577,4	8,1	365,1	8,4	418,1	9,4	0,6	0,5	0,6	0,2	2,2	0,2	1,7	0,1	2,2	0,6
L3	48,5	9,1	58,8	8,2	545,0	8,3	345,3	8,1	395,8	9,1	0,7	0,7	0,7	0,2	2,6	0,2	2,0	0,2	2,6	0,6
L4	49,7	9,4	61,6	8,5	574,5	8,4	364,5	8,2	416,6	14,3	0,6	0,6	0,7	0,2	2,2	0,2	1,8	0,2	2,3	0,5
L5	52,2	9,3	62,6	8,2	575,9	8,6	364,4	8,6	415,2	9,3	0,7	0,6	0,7	0,2	2,2	0,2	1,8	0,2	2,3	0,7
L6	51,1	9,2	62,2	8,7	574,4	8,3	364,5	8,2	416,1	10,6	0,6	0,5	0,6	0,2	2,0	0,2	1,7	0,1	2,2	0,4
L7	51,5	9,8	62,2	8,4	574,4	9,2	365,4	8,3	418,1	9,2	0,6	0,5	0,7	0,2	2,0	0,2	1,6	0,2	2,1	0,5
L8	51,6	9,2	62,0	8,8	574,5	8,4	365,2	8,6	416,5	9,0	0,6	0,5	0,7	0,2	2,2	0,2	1,8	0,2	2,2	0,5
L9	49,1	9,6	59,0	8,2	548,3	8,7	347,0	8,1	394,2	9,2	0,6	0,5	0,7	0,1	2,2	0,2	1,8	0,2	2,3	0,4
PROM	50,7	9,5	61,5	8,4	569,8	8,5	360,5	8,4	410,3	9,9	0,6	0,6	0,7	0,2	2,2	0,2	1,8	0,2	2,3	0,5

Tabla 4.7: Experimento 2: Comparación entre TPR - TNR- ACC (1.754 muestras)

	ALX		SQZ		IV3		RN5		VGG	
	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM
TPR	61,11 %	65,56 %	28,89 %	60,00 %	47,78 %	58,89 %	58,89 %	71,11 %	68,89 %	62,22 %
TNR	97,86 %	93,57 %	93,57 %	91,43 %	80,71 %	80,71 %	95,71 %	88,57 %	87,14 %	90,00 %
ACC	83,48 %	82,61 %	68,26 %	79,13 %	67,83 %	72,17 %	81,30 %	81,74 %	80,00 %	79,13 %

4.1.3. Experimemto 3: Espectrogramas en color de fragmentos de sonidos

Para este experimento se probó, como técnica de aumentación de datos, dividir los sonidos en segmentos de 1 segundo con un solapamiento del 50 %, generando para cada sonido obtenido espectrogramas en las 13 escalas de colores consideradas, obteniéndose un total de 15.184 muestras.

Al igual que los experimentos anteriores, en la Tabla 4.8 se puede observar que el accuracy obtenido para los dos tipos de redes tiene valores similares, alcanzando un máximo del 95.65 %.

Tabla 4.8: Experimento 3: Accuracy (15.184 muestras)

	ALX		SQZ		IV3		RN5		VGG	
	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM
Lote0	91.30 %	82.61 %	69.57 %	78.26 %	82.61 %	69.57 %	78.26 %	82.61 %	73.91 %	82.61 %
Lote1	78.26 %	78.26 %	69.57 %	73.91 %	78.26 %	69.57 %	78.26 %	73.91 %	86.96 %	73.91 %
Lote2	82.61 %	86.96 %	82.61 %	73.91 %	69.57 %	73.91 %	82.61 %	86.96 %	91.30 %	82.61 %
Lote3	73.91 %	78.26 %	56.52 %	69.57 %	73.91 %	65.22 %	65.22 %	73.91 %	78.26 %	73.91 %
Lote4	78.26 %	69.57 %	69.57 %	73.91 %	73.91 %	73.91 %	82.61 %	78.26 %	86.96 %	78.26 %
Lote5	82.61 %	78.26 %	78.26 %	78.26 %	73.91 %	65.22 %	82.61 %	82.61 %	91.30 %	73.91 %
Lote6	95.65 %	91.30 %	82.61 %	86.96 %	91.30 %	86.96 %	91.30 %	95.65 %	91.30 %	95.65 %
Lote7	91.30 %	78.26 %	73.91 %	82.61 %	86.96 %	73.91 %	82.61 %	86.96 %	69.57 %	73.91 %
Lote8	78.26 %	73.91 %	78.26 %	78.26 %	69.57 %	65.22 %	69.57 %	73.91 %	78.26 %	78.26 %
Lote9	86.96 %	82.61 %	73.91 %	73.91 %	86.96 %	60.87 %	91.30 %	82.61 %	82.61 %	78.26 %
MAX	95.65 %	**91.30 %**	82.61 %	86.96 %	91.30 %	86.96 %	**91.30 %**	95.65 %	91.30 %	95.65 %
MIN	73.91 %	69.57 %	56.52 %	69.57 %	69.57 %	60.87 %	65.22 %	73.91 %	69.57 %	73.91 %
PROM	83.91 %	80.00 %	73.48 %	76.96 %	78.70 %	70.43 %	80.43 %	81.74 %	83.04 %	79.13 %
RANGO	21.74 %	21.74 %	26.09 %	17.39 %	21.74 %	26.09 %	26.09 %	21.74 %	21.74 %	21.74 %

La Tabla 4.9 muestra que las ELM tienen un tiempo de entrenamiento mucho menor que varía entre 17.2 y 22.9 seg, mientras que en las CNN varia entre 455.4 y 5.103,5 seg. Los tiempos de entrenamiento en las ELM no llegan a los 2 segundos, mientras que en las CNN se acercan en el peor caso a 17 segundos.

Tabla 4.9: Experimento 3: Tiempos de entrenamiento y test (15.184 muestras)

Lote	TIEMPO DE ENTRENAMIENTO										TIEMPO DE TEST									
	ALX		SQZ		IV3		RN5		VGG		ALX		SQZ		IV3		RN5		VGG	
	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM
L0	471,2	25,5	551,7	17,8	5.232,8	17,0	3.231,8	18,0	3.621,0	22,1	3,5	1,4	2,9	0,7	14,9	0,7	12,1	0,7	14,1	1,3
L1	474,7	24,2	568,2	17,7	5.354,6	21,7	3.375,5	18,2	3.943,8	20,9	3,2	1,4	3,0	0,7	14,4	0,8	12,1	0,7	14,3	1,3
L2	426,2	21,8	504,4	16,8	4.812,6	15,7	3.028,3	15,7	3.534,9	19,8	4,3	1,7	4,0	1,0	19,4	1,0	16,3	0,9	18,8	1,7
L3	440,3	20,7	505,8	16,0	4.886,7	17,3	3.056,1	16,0	3.580,4	23,3	4,3	1,7	3,8	0,9	19,4	1,0	16,6	0,9	19,2	1,6
L4	440,4	20,8	528,0	20,5	4.997,7	16,8	3.149,8	16,9	3.670,6	24,3	4,1	1,6	3,9	0,9	18,0	0,9	15,2	0,9	17,7	1,7
L5	470,3	21,1	544,9	16,9	5.119,0	17,8	3.225,2	17,5	3.797,0	23,1	3,5	1,4	3,1	0,8	15,5	0,8	13,0	0,8	15,3	1,5
L6	476,6	21,6	574,4	18,2	5.479,8	17,7	3.458,5	18,3	4.032,0	22,5	3,4	1,4	3,1	0,7	15,0	0,8	12,6	0,7	14,7	1,4
L7	431,1	20,3	512,5	17,3	4.839,2	16,1	3.048,2	17,1	3.610,5	27,1	3,8	1,6	3,6	0,9	17,2	0,9	14,5	0,9	17,0	1,5
L8	473,0	20,8	562,2	20,9	5.207,9	16,3	3.321,2	17,2	3.895,6	21,0	3,6	1,4	3,2	0,8	15,3	0,8	12,9	0,8	15,4	1,4
L9	449,9	24,4	542,9	17,0	5.104,8	17,6	3.223,1	16,9	3.782,6	24,3	4,3	1,6	3,9	0,9	19,1	0,9	16,2	0,9	19,2	1,7
PROM	455,4	22,1	539,5	17,9	5.103,5	17,4	3.211,8	17,2	3.746,9	22,9	3,8	1,5	3,4	0,8	16,8	0,8	14,2	0,8	16,6	1,5

En la Tabla 4.10 vemos que los valores de TPR, TNR y ACC, son similares para las CNN y ELM, en todas las arquitecturas consideradas. Comparada con el experimento 1 y 2, se ve un aumento significativo del TPR alcanzando un máximo del 88.57 % para las CNN y 87.14 % para las ELM. El TNR disminuye levemente sus valores en relación al experimento 2, pero igualmente mantiene valores aceptables de clasificación.

Tabla 4.10: Experimento 3: Comparación entre TPR - TNR- ACC (15.184 muestras)

	ALX		SQZ		IV3		RN5		VGG	
	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM	CNN	ELM
TPR	88,57 %	83,57 %	82,86 %	87,14 %	85,71 %	76,43 %	88,57 %	87,14 %	87,86 %	85,71 %
TNR	82,62 %	79,01 %	70,87 %	74,13 %	76,75 %	68,77 %	78,17 %	80,24 %	81,71 %	77,30 %
ACC	83,91 %	80,00 %	73,48 %	76,96 %	78,70 %	70,43 %	80,43 %	81,74 %	83,04 %	79,13 %

4.1.4. Experimento 4: Espectrogramas en color de sonido original y fragmentos

Con el objetivo de estudiar el comportamiento de las ELM al aumentar la cantidad de muestras, en este último experimento se creó una base de datos

uniendo los espectrogramas de los sonidos originales a los espectrogramas de los fragmentos, obteniéndose un total de 16.939 muestras.

Como muestra la Tabla 4.11, no se observa una variación significativa en el accuracy, manteniéndose valores similares entre las dos tipos de redes.

Tabla 4.11: Experimento 4: Accuracy (16.939 muestras)

	ALX		**SQZ**		**IV3**		**RN5**		**VGG**	
	CNN	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**
Lote0	78.26%	78.26%	82.61%	75.00%	78.26%	85.71%	69.57%	85.71%	78.26%	82.14%
Lote1	69.57%	73.91%	78.26%	69.23%	78.26%	69.23%	69.57%	73.08%	78.26%	69.23%
Lote2	78.26%	69.57%	86.96%	84.62%	65.22%	57.69%	78.26%	65.38%	86.96%	69.23%
Lote3	60.87%	86.96%	52.17%	75.00%	69.57%	62.50%	60.87%	81.25%	73.91%	84.38%
Lote4	73.91%	73.91%	73.91%	71.43%	73.91%	67.86%	69.57%	82.14%	82.61%	78.57%
Lote5	82.61%	86.96%	82.61%	71.43%	82.61%	64.29%	82.61%	85.71%	91.30%	78.57%
Lote6	95.65%	82.61%	78.26%	73.08%	95.65%	65.38%	91.30%	76.92%	91.30%	84.62%
Lote7	86.96%	91.30%	56.52%	76.00%	82.61%	68.00%	82.61%	80.00%	78.26%	88.00%
Lote8	82.61%	82.61%	69.57%	70.37%	69.57%	81.48%	69.57%	88.89%	82.61%	70.37%
Lote9	78.26%	73.91%	78.26%	89.29%	82.61%	71.43%	91.30%	78.57%	78.26%	82.14%
MAX	95.65%	**91.30%**	86.96%	89.29%	95.65%	85.71%	**91.30%**	88.89%	91.30%	88.00%
MIN	60.87%	69.57%	52.17%	69.23%	65.22%	57.69%	60.87%	65.38%	73.91%	69.23%
PROM	78.70%	80.00%	73.91%	75.54%	77.83%	69.36%	76.52%	79.77%	82.17%	78.73%
RANGO	34.78%	21.74%	34.78%	20.05%	30.43%	28.02%	30.43%	23.50%	17.39%	18.77%

La Tabla 4.12 muestra que la diferencia de los tiempos de entrenamiento entre las CNN y ELM continúan creciendo al aumentar el número de muestras. Lo mismo sucede con los tiempos de entrenamiento, y sus valores continúan siendo despreciables en comparación con los tiempos de entrenamiento. En Las ELM, el tiempo de entrenamiento varían entre 17,9 y 28,8 seg y para las CNN varia entre 548,2 y 5,857,4 seg.

Tabla 4.12: Experimento 4: Tiempo de entrenamiento y test (16.939 muestras)

Lote	**TIEMPO DE ENTRENAMIENTO**										**TIEMPO DE TEST**									
	ALX		**SQZ**		**IV3**		**RN5**		**VGG**		**ALX**		**SQZ**		**IV3**		**RN5**		**VGG**	
	CNN	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**
L0	561,5	31,7	674,7	17,7	6.077,2	17,9	3.877,8	17,9	4.309,2	24,5	3,6	1,7	3,4	0,9	17,3	0,9	13,9	0,9	16,2	1,5
L1	557,8	37,7	658,5	18,3	6.056,0	29,6	4.224,2	17,7	4.595,4	23,6	3,6	1,5	3,4	0,8	16,2	0,8	14,9	0,9	16,4	1,5
L2	523,8	40,5	622,9	18,3	5.567,4	31,0	4.032,8	17,1	4.270,7	21,3	4,9	3,0	4,4	1,3	21,6	1,6	19,1	1,3	21,5	2,0
L3	534,6	38,5	613,0	16,9	5.623,1	28,6	4.000,6	18,9	4.315,0	21,4	4,9	2,6	4,7	1,1	21,9	1,3	20,0	1,1	21,8	2,9
L4	530,3	21,9	634,2	18,1	5.725,4	17,2	3.836,3	18,2	4.351,9	21,8	4,5	2,8	4,2	1,2	20,4	1,2	17,0	1,6	19,9	2,9
L5	571,5	27,0	650,5	18,8	5.927,7	17,7	3.807,9	18,6	4.478,6	22,1	4,0	1,6	3,9	0,9	17,4	0,9	15,3	0,9	17,3	1,6
L6	565,3	23,0	644,9	19,2	6.147,3	18,5	3.943,9	17,9	4.616,2	25,0	3,8	1,5	3,4	0,8	16,8	0,8	14,2	0,9	16,6	1,5
L7	539,3	22,1	647,9	18,6	5.783,8	17,9	3.701,4	17,5	4.365,5	25,8	4,3	1,7	4,3	1,0	19,2	1,0	16,0	1,0	19,0	1,8
L8	569,3	22,6	645,0	21,4	5.936,9	18,8	3.827,2	17,4	4.459,6	27,8	4,0	1,6	3,6	0,9	17,2	1,0	15,0	0,9	17,0	1,6
L9	528,5	23,2	614,0	17,4	5.729,5	17,2	3.837,3	18,4	4.299,0	22,3	4,8	1,8	4,4	1,2	21,5	1,3	18,3	0,9	20,8	2,8
PROM	548,2	28,8	640,5	18,5	5.857,4	21,4	3.908,9	17,9	4.406,1	23,6	4,2	2,0	4,0	1,0	18,9	1,1	16,4	1,0	18,6	2,0

En la Tabla 4.13 vemos que los valores de TPR, TNR y ACC son similares para las CNN y ELM, en todas las arquitecturas consideradas. Comparada con el experimento 3, resultan valores inferiores.

Tabla 4.13: Experimento 4: Comparación entre TPR - TNR- ACC (16.939 muestras)

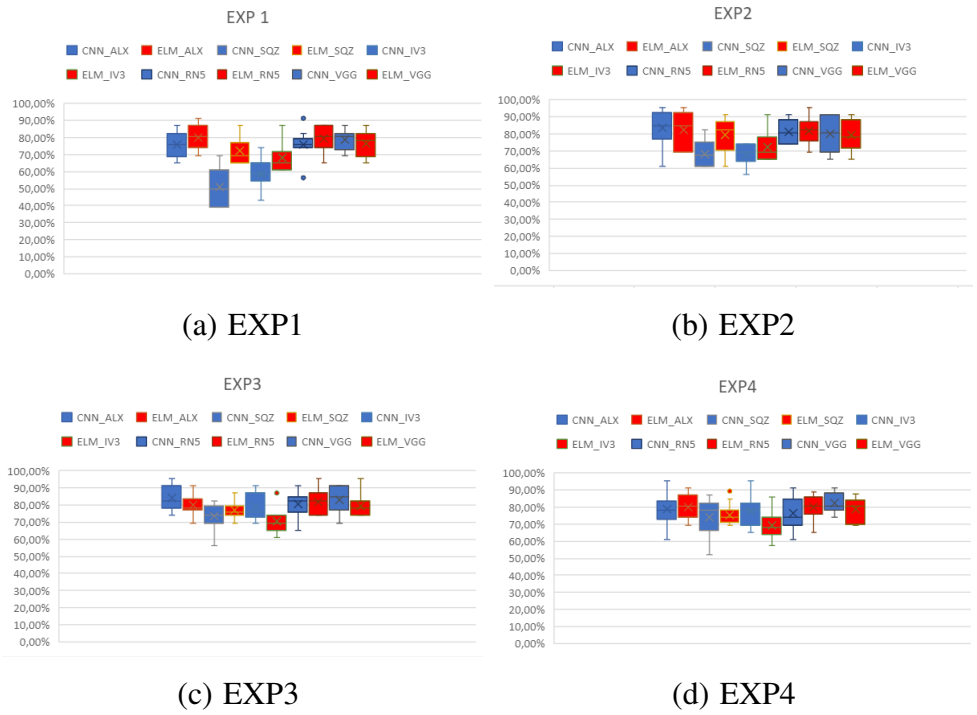
	ALX		**SQZ**		**IV3**		**RN5**		**VGG**	
	CNN	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**	**CNN**	**ELM**
TPR	64,44 %	66,67 %	64,44 %	57,78 %	64,44 %	57,78 %	60,00 %	67,78 %	73,33 %	65,56 %
TNR	87,86 %	83,57 %	80,00 %	82,14 %	86,43 %	74,29 %	87,14 %	87,14 %	87,86 %	82,86 %
ACC	78,70 %	80,00 %	73,91 %	75,54 %	77,83 %	69,36 %	76,52 %	79,77 %	82,17 %	78,73 %

4.1.5. Análisis de los resultados de los EXP1, EXP2, EXP3 Y EXP4

Una revisión completa (202 artículos citados), y actual (publicada año 2022), sobre la detección de la enfermedad de Parkinson, utilizando análisis del habla y la voz [153] en base a todo tipo de técnicas de machine learning (desde clasificadores clásicos hasta CNN) arroja que, la mayoría de las bases de datos presentadas oscilan entre las 50 y 100 personas, y los resultados de performance están entre el 77 % y 99 %. En el caso de este trabajo, la base de datos considerada tiene 135 personas y se logra una performance oscilando en 83,91 % para la CNN, y 81,74 % para la ELM. Lo anterior indica que, tanto en numero de personas analizadas como en la performance, nuestros resultados son comparables con la literatura.

Para realizar un análisis de performance se presentan las Figura 41 y 42. La Figura 41 corresponde a diagramas de caja por experimento y la Figura 42 corresponde a diagramas de caja por arquitectura. Las cajas azules corresponden a las arquitecturas CNN y las cajas rojas a las ELM.

Considerando la Figura 41 se observa que los experimentos con mayor cantidad de muestras tienen los mejores resultados (EXP3 Y EXP4). Es decir que tienen la más baja dispersión de los datos y el promedio de performance más elevado. En particular, en el EXP3, el cual presenta los mejores resultados, en general, la dispersión para el clasificador ELM es menor que la dispersión para el clasificador CNN, lo que muestra un buen comportamiento de la propuesta del articulo.

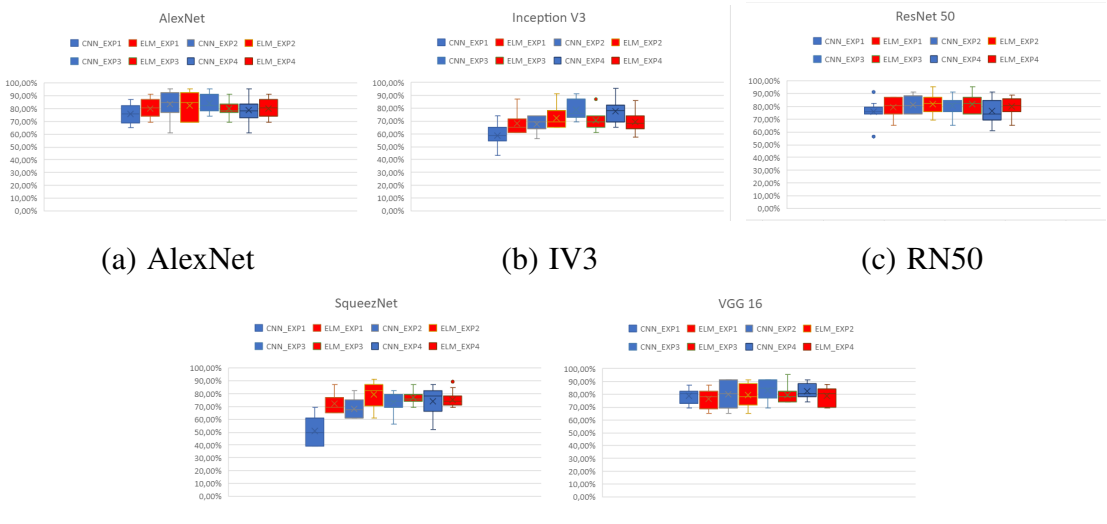


(a) EXP1 (b) EXP2

(c) EXP3 (d) EXP4

Figura 41: Diagrama de cajas por experimento

En la Figura 42, se observa que la red AlexNet (Figura 42(a)) tiene el mejor equilibrio entre dispersión y el valor medio de performance. En particular, en el EXP3, se observa una menor dispersión de la ELM respecto de la CNN.



(a) AlexNet (b) IV3 (c) RN50

(d) SQZ (e) VGG16

Figura 42: Diagrama de cajas por Arquitectura

Las Tablas 4.3, 4.6, 4.9 y 4.12 presentan los tiempos de entrenamiento y test para los EXP1, EXP2, EXP3 y EXP4 respectivamente. Es posible mencionar que el tiempo de entremiento va aumentando a medida que existe una mayor cantidad de muestras para las CNN y ELM. No obstante, el tiempo de entrenamiento de las ELM es notoriamente menor, lo cual confirma la hipótesis principal de este trabajo. Respecto de los tiempos de testing, estos son sustancialmente menores que los tiempo de entrenamiento. En general, el tiempo de testing es similar para las CNN y las ELM en los dos experimentos de menor cantidad de muestras (EXP1 y EXP2), para los experimentos de mayor cantidad de muestras (EXP3 y EXP4), el tiempo de test es sustancialmente menor para las ELM.

4.1.6. Análisis comparativo entre CNN y Modelos Logísticos

En este apartado se comparan dos técnicas para la clasificación en enfermos y no enfermos, es decir, personas con y sin EP respectivamente.

La primera técnica propuesta en esta tesis realiza la clasificación mediante el uso de espectrogramas generados a partir de las señales de voz, procesados con CNN. La segunda técnica utiliza las medidas de disfonía extraídas y seleccionadas de las señales de voz, propuestas en [154] para realizar la clasificación binaria. Esta última toma una selección de 11 variables utilizando medidas acústicas de la señal de audio, y clasifica los audios con modelos logísticos.

Para realizar la comparación de las dos metodologías de trabajo, los experimentos se realizaron sobre las mismas muestras de la base de datos. Se utilizaron 135 fonaciones de la vocal /a/ sostenida correspondiente a 118 individuos, donde se descartaron aquellas con mayores alteraciones, y se tomó uno, dos o tres audios por persona, según disponibilidad. En todos los casos, se seleccionaron submuestras de entrenamiento y test teniendo en cuenta a los individuos, de manera que los audios de una misma persona solo quedaran en entrenamiento, o en test, con todas sus repeticiones. En el caso de las CNN se consideraron tres conjuntos disjuntos: entrenamiento, validación y test.

4.1.6.1. Variables acústicas para la clasificación con modelo logístico

Giuliano y otros en [154], analizan varias medidas de disfonías con el fin de obtener un sistema mínimo de variables. Se realiza el análisis de la señal de voz con el Voice Analysis Toolbox (VAT) [108, 155, 156]. EL VAT está compuesto por una serie de rutinas en MATLAB optimizadas para la pronunciación de la letra /a/. El sistema entrega un total de 339 medidas que se clasificaron en 4 grupos según los problemas más comunes que presentan las voces disfuncionales de EP. Los autores presentan un análisis estadístico, y seleccionan por su relevancia, 11 variables del conjunto de las 339 variables.

Las variables seleccionadas están indicadas en la Tabla 4.14 y en los párrafos siguientes se hace una descripción de los grupos mencionados.

Tabla 4.14: Clasificación de las 11 medidas de disfonía utilizadas según grupo de pertenencia (la variable proporción indica Cantidad Seleccionadas / total del grupo).

Grupo	Medidas	Variables	Proporción
G1	Desvío de la periodicidad	V1-V51	2/209
G2	Variaciones de la amplitud	V34	1/22
G3	Ruido	V59, V70, V338	3/24
G4	Problemas en la articulación	V137, V141, V152, V71, V75	5 /84

La señal asociada a la fonación de una vocal se puede modelar como una señal periódica, caracterizada por su frecuencia fundamental. Durante la fonación sostenida de una vocal, el desvío de la periodicidad es más notable en las voces patológicas que en las voces no enfermas. Para el estudio de este desvío, VAT entrega las medidas típicas de las variaciones de la curva de F0, junto con medidas basadas en análisis no lineal de señales GQ (*Glottal Quotient*), PPE (*Pitch Period Entropy*), RPDE (*Recurrence Period Density Entropy*), parámetros derivados de los coeficientes de aproximación y detalles del Análisis Wavelets de la curva de F0. Estas medidas corresponden el grupo G1.

Otro problema que surge, en el análisis de voces patológicas, es la variación de la amplitud de señal. Para la cuantificación de estas variaciones se utilizan los parámetros del grupo G2 calculados con VAT.

El cierre incompleto de las cuerdas vocales produce un flujo de aire turbulento que se manifiesta como ruido (G3: ruido). Para el análisis de este fenómeno que en las voces patológicas se presenta con mayor intensidad, VAT entrega los parámetros estándar junto con una serie de medidas más modernas. Las medidas son: HNR (*Harmonics to Noise Ratio*), NHR (*Noise to Harmonics Ratio*), GNE (*Glottal to noise excitation*), VFER (*Vocal fold excitation ratio*), EMD-ER (*Empirical mode decomposition excitation ratio*), y DFA (*Detrended Fluctuation Analysis*).

Los enfermos de Parkinson, por lo general, tienen problemas en la articulación necesaria para la fonación. Esto puede medirse por intermedio de parámetros derivados de los Coeficientes Cepstrales MFCCs (*Mel Frequency Cepstral Coefficients*), que caracterizan al tracto vocal e integran el grupo G4.

Resumiendo, en el trabajo mencionado anteriormente [154], se tomaron muestras de la fonación de la /a/ de 108 individuos, y se analizaron las 339 variables indicadas en la Tabla 4.14. Luego se seleccionaron las 11 variables que mejor representan al conjunto. De las once variables que resultaron mejor jerarquizadas, 2 pertenecen al grupo G1, 1 al grupo G2, 3 son del grupo G3, y 5 al grupo G4. En el conjunto total de variables seleccionadas, se destaca la presencia de variables en cada uno de los cuatro grupos teóricos utilizados como referencia.

Los mismos investigadores realizaron luego un análisis predictivo de regresión logística, tomando como base las 11 variables calculadas en la nueva muestra de 135 audios. A partir de estas, se aplicó una selección automática de variables utilizando el método de regresión logística por pasos (*stepwise*) [157], de modo de considerar un modelo con variables significativas para la predicción. Se dividió la muestra en entrenamiento y test para validación y se calculó la matriz de confusión. Por último, las métricas de evaluación se calcularon a partir de la matriz de confusión. A partir de las 11 variables, sumadas a género y edad, se realizó una regresión logística por pasos. El modelo resultante deja seleccionadas sólo 5 variables: tsV034, tsV059, tsV075, tsV141, y tsV338.

En la Tabla 4.15 se muestran los coeficientes estimados del modelo, el que obtuvo un valor de AIC = 54,686 y una capacidad predictiva, medida a través del área bajo la curva ROC, igual a AUC = 0,91148. Considerando un punto de corte de 0,5 para la clasificación, la exactitud obtenida es accuracy = 85.36 %.

Tabla 4.15: Coeficientes estimados en el Modelo Logístico

Variable	**Std.**	**Error**	**z**	**valor**
Constante	-16,55	5,31	-3,12	0,002
V034	26,62	15,85	1,68	0,093
V059	1011,73	363,59	2,78	0,005
V075	1,25	0,49	2,55	0,011
V141	351,18	98,30	3,57	0,000
V338	13,01	7,31	1,78	0,075

4.1.6.2. Metodología de Comparación

Se consideró la matriz de confusión para evaluar los algoritmos utilizados para la clasificación de los audios de test. Se consideran la condición de enfermo (EP) y no enfermo (NEP) de Parkinson versus la condición de clasificación o predicción utilizando cada uno de los dos métodos (valores predichos).

Se calcularon luego indicadores usuales, los que se resumen en la matriz de confusión dada en la Tabla 4.16. En esta se presenta la siguiente notación: TP corresponde a los verdaderos positivos (EP predichos como EP); FN a falsos negativos (EP predichos como NEP); TN corresponde a verdaderos negativos (NEP predichos como NEP) y por último FP indicando falsos positivos (NEP predichos como EP).

Tabla 4.16: Matriz de confusión

	EP predicción	**NEP predicción**	
EP reales	TP	FN	TEP
NEP reales	FP	TN	TNEP
	TEPP	TNEP	Total

Para los totales, se tiene TEP al total de EP y TNEP es total de NEP reales; mientras que TEPP es el total de EP y TNEP es el total de NEP predichos.

A continuación, se definen los indicadores utilizados y calculados a partir de la matriz de confusión:

- ACC representa la Exactitud (*Accuracy*) o proporción de datos correctamente clasificados.
- Tasa de Error es la proporción del total de datos mal clasificados.
- Sensibilidad (*Recall -Hit Rate*), o TPR (*True Positive Rate*), representa la tasa de verdaderos positivos entre el total de enfermos, mide la exhaustividad, los datos correctamente clasificados del grupo de interés, es decir que tan precisa es la prueba de detección para identificar la enfermedad entre las personas que la padecen.
- Especificidad, o TNR (*True Negative Rate*), representa la tasa de verdaderos negativos que mide la selectividad, es decir, la tasa de negativos que se identifican correctamente, los no enfermos bien detectados.
- Exactitud balanceada (*Balanced Accuracy*), es el promedio de la tasa de verdaderos positivos (TPR) y verdaderos negativos (TNR).
- FNR (*False Negative Rate*), es la tasa de errores o tasa de Falsos Negativos. Representa la proporción de mal predichos como no enfermos en el total de enfermos.
- FPR (*False Positive Rate*), es la tasa de Falsos Positivos, es decir, la proporción de no enfermos mal predichos como enfermos en el total de no enfermos.
- PPV (*Positive Predictive Value*), es la Precisión o Valor Predictivo Positivo. Indica la proporción de sujetos con detección positivas que realmente tienen la enfermedad, es decir, enfermos detectados del total de los clasificados como enfermos.
- NVP (*Negative Predictive Value*), es el Valor Predictivo Negativo. Indica la proporción de detecciones negativas que realmente no tienen la enfermedad, es decir, los predichos como no enfermos del total de no enfermos.

- FDR (*False Discovery Rate*), es la Tasa de Descubrimiento Falso, indica la razón de mal predicho como enfermos del total de los predichos como enfermos.

  FDR = FP / (FP + TP) = 1 – PPV

- FOR (*False Omission Rate*) es la tasa de Falsas Omisiones, indica la razón de mal predichos como no enfermos del total de los predichos como no enfermos.

  FOR = FN / (FN + TN) = 1 – NVP.

- F1 Score es una medida de la exactitud. Se calcula a partir de la precisión y la sensibilidad usando la media armónica de estas. Si los datos están desbalanceados, es una mejor medida que el Accuracy.

  F1 = 2 * [2 TP / (2 * TP + FP + FN)] = (PPV + TPR) / (PPV + RPR)

- MCC (*Matthews Correlation Coefficient*), es el Coeficiente de Correlacion de Matthews tiene en cuenta los positivos y negativos verdaderos y falsos. En general, se considera una medida equilibrada que se puede utilizar incluso si las clases son de tamaños muy diferentes (-1 Muy malo, 0 Malo, 1 Muy bueno).

- LR (*Likelihood Ratio*), conocido en español como razón de verosimilitud, se define como la razón entre la posibilidad de observar un ser clasificado enfermo en los pacientes con EP versus la posibilidad de ser clasificado como enfermo en personas sin la patología.

  LR+ = (sensibilidad/1-especificidad)

  LR- = (1-sensibilidad / especificidad).

  El indicador LR+ es la chance de ser realmente EP versus ser no enfermo, si fue clasificado como enfermo.

  El indicador LR- es la chance de ser realmente EP versus ser no enfermo, si fue clasificado como no enfermo.

- OR (*Odds Ratio*), es el cociente de la razón de verosimilitud positiva respecto de la negativa (OR = LR+ / LR-). Este índice representa la probabilidad relativa de que una persona enferma sea clasificada como tal en comparación con la probabilidad de que una persona no enferma sea incorrectamente clasificada como enferma.

Tabla 4.17: Indicadores para la evaluación de la clasificación

Indicador	**Fórmula**
ACC Accuracy o Exactitud	(TP + TN)/ TOTAL
Tasa de Error	(FP + FN) / TOTAL = 1 - ACCURACY
TPR Sensibilidad o tasa de verdaderos positivos	TPR = TP / (TP + FN) = 1 - FNR
TNR Especificidad Tasa de Verdaderos Negativos	TNR = TN / (TN + FP) = 1 - FPR
Exactitud Balanceada o Balanced Accuracy	(TPR + TNR) / 2
FNR: Tasa de Errores o Tasa De Falsos Negativos	FNR = FN / (FN + TP) = 1 - TPR
FPR Tasa de Caída o Tasa de Falso Positivos	FPR = FP / (FP + TN) = 1 - TNR
PPV Precisión o Valor Predictivo Positivo	PPV = TP / (TP + FP) = 1 - FDR
NVP Valor Predictivo Negativo	NVP = TN / (TN + FN) = 1 - FOR
FDR Tasa de Descubrimiento Falso	FDR = FP / (FP + TP) = 1 - PPV
FOR Tasa de Falsas Omisiones	FOR = FN / (FN + TN) = 1 - NVP
F1 Score	F1 = 2 * [2 TP / (2 * TP + FP + FN)] = (PPV + TPR) / (PPV + RPR)
Coeficiente de Correlación de Matthews	MCC = [(TP * TN) - (FP * FN)] / SQRT ((TP + FP) * (TP + FN) * (TN + FP) * (TN + FN))
Razón de verosimilitud positiva (LR+)	LR+ =TPR / (1-TNR)= TPR /FPR
Razón de verosimilitud negativa (LR-)	LR- = (1-TPR) / TNR= FNR / TNR
Odds Ratio (OR)	OR=LR+ / LR- == (TP/FN) / (FP/TN)

4.1.6.3. Comparación de indicadores

En la Tabla 4.18 se observa los valores de los indicadores estudiados según el modelo logístico con las variables acústicas y el modelo CNN ALEXNET con las imágenes.

La exactitud de la clasificación resulta 85 % con las medidas acústicas mientras que 88 % con las imágenes. Son valores similares un poco mejor en el caso de las imágenes. Esto también se refleja en las tasas de error que son de 15 % y 12 %, en correspondencia con la exactitud.

Tabla 4.18: Índices obtenidos en la evaluación de la clasificación según el modelo logístico y el modelo CNN Alexnet

Indicador	**Modelo Logístico**	**CNN ALEXNET**
ACC exactitud	0,85	0,88
Tasa de Error	0,15	0,12
TPR Sensibilidad	0,79	0,87
TNR Especificidad Tasa de Verdaderos Negativos	0,91	0,89
Exactitud Balanceada	0,85	0,88
PPV Precisión o Valor Predictivo Positivo	0,88	0,88
NVP Valor Predictivo Negativo	0,83	0,88
F1 Score	0,83	0,87
Coeficiente de Correlación de Matthews	0,71	0,76
FDR Tasa de Descubrimiento Falso	0,21	0,13
FOR Tasa de Falsas Omisiones	0,09	0,11
Razón de verosimilitud positiva (LR+)	8,68	8,21
Razón de verosimilitud negativa (LR-)	0,23	0,15
Odds Ratio (OR)	37,50	55,28

La sensibilidad es peor en modelos logísticos que la clasificación con CNN (79 % y 87 %), lo que indicaría que la tasa de enfermos bien clasificados es mejor en CNN. En cambio, la especificidad es mejor levemente en el modelo logístico (91 % y 89 %), lo que indica mejor precisión en los verdaderos no enfermos clasificados como tal.

F1 Score resultó similar, levemente mayor para la clasificación de imágenes (87 %) respecto la clasificación de medidas acústicas (83 %)

El MCC indica correlación positiva alta de 71 % y 76 %, un poco mejor para imágenes. La razón de verosimilitud da en los dos casos valores para LR+ y LR- correspondiente a la categoría de buena utilidad de la clasificación (LR+ entre 5 y 10 y LR- entre 0,1 y 0,2).

Siendo el OR mejor en el caso de clasificación de imágenes. Hay 55 veces más chance de clasificar a un enfermo como tal (respecto a clasificarlo mal), que de clasificar como enfermo a un no enfermo de Parkinson (respecto a clasificarlo bien).

4.1.6.4. Discusión

Los resultados se compararon utilizando diferentes indicadores (sensibilidad, especificidad, exactitud, tasa de error, caída de falsos positivos, precisión, tasa de descubrimiento falso, tasa de falsas omisiones, F1 score, Coeficiente de Correlación de Matthew, razón de verosimilitud, *odds ratio*). Los resultados muestran rendimientos similares en los indicadores utilizados y se destaca un mayor *odds ratio* en la clasificación utilizando espectrogramas generados a partir de las señales de voz.

La metodología que utiliza los parámetros acústicos seleccionados tiene como ventaja que las variables tienen un correlato físico de afecciones en el sistema fonatorio. Sin embargo, tiene la desventaja que se trabaja con muchas variables. Se necesita realizar la selección de las mismas y su cálculo debería realizarse siempre de la misma manera.

Las redes neuronales convolucionales, en comparación con los algoritmos tradicionales de ML, aprenden características de forma automática. Ésta particularidad de las CNN, evita el proceso previo de encontrar descriptores, que serían necesarios para entrenar otro tipo de clasificadores. Esta es una de sus mayores fortalezas ya que reduce el tiempo de preprocesamiento de las muestras. Sin embargo, para entrenar el modelo se necesita un gran número de muestras etiquetadas.

Capítulo 5

Conclusiones y prospectivas

5.1. Conclusiones

En esta tesis se estudiaron distintos métodos de aprendizaje profundo que utilizan espectrogramas de señales de voz para la clasificación de enfermos y no enfermos de Parkinson. Los sistemas de diagnóstico basados en el análisis de la voz presentan ventajas pues permiten el diagnóstico no invasivo y precoz de la enfermedad. Se creó una base de datos de espectrogramas de las señales de voz de la vocal /a/ sostenida a partir de las muestras del repositorio de grabaciones de enfermos y no enfermos de Parkinson disponible en el repositorio de la UNLaM. Se propusieron dos estrategias de aumentación de datos para el repositorio de espectrogramas de los sonidos originales. La primera estrategia consistió en crear espectrogramas a partir de la señal de voz original considerando distintas paletas de colores. La segunda consistió en fragmentar el sonido original en segmentos de 1 segundo con el 50 % de solapamiento. Para cada uno de estos fragmentos se generaron los espectrogramas en las 13 paletas de colores de MatLab seleccionadas.

Se probaron diferentes arquitecturas de CNN, aplicadas al repositorio ampliado con la primera estrategia de aumentación de datos. Se aplicó la técnica de transferencia de aprendizaje debido a la cantidad de muestras disponibles. Para todas las arquitecturas, se observó una mejora en los indicadores de performan-

ce para el conjunto de datos aumentado. Los niveles de performance alcanzados muestran que la estrategia de aumentación de datos a través de la utilización de diferentes paletas de colores, es pertinente y que las CNN resuelven el problema con niveles de precisión aceptables.

Se realizó una comparación entre la metodología utilizada en esta tesis, CNNs aplicadas a los espectrogramas obtenidos de las señales de voz, y el análisis acústico realizado por otros autores sobre las mismas muestras de datos. De esta forma, se pudo comprobar que los resultados obtenidos fueron similares, con una leve mejora para la técnica de las CNNs, que además presenta la ventaja de no requerir un pre - procesamiento de las señales de audio para encontrar los mejores descriptores de las muestras analizadas.

Para comprobar la hipótesis planteada en esta tesis, se realizaron experimentos sobre modelos de CNN y ELM midiendo los tiempos de entrenamiento, validación y test, así como también se registraron las medidas de rendimiento de cada modelo: accuracy, eficiencia y especificidad. Se realizaron 4 experimentos sobre diferentes conjuntos de espectrogramas:

- Experimento 1: espectrogramas en escala de grises de los sonidos originales.
- Experimento 2: espectrogramas en las 13 paletas de colores seleccionadas aplicadas a los sonidos originales.
- Experimento 3: espectrogramas en las 13 paletas de colores de los segmentos de 1 segundo.
- Experimento 4: conjunto de espectrogramas del experimento 2, más los del experimento 3.

En cada experimento se aplicaron modelos de CNN y ELM sobre 5 arquitecturas consideradas. Las 5 arquitecturas de Redes Neuronales Convolucionales pre-entrenadas seleccionadas fueron: AlexNet, VGG-16, SqueezeNet, Inception V3, y ResNet-50.

Para reducir el tiempo de entrenamiento, en esta tesis se presentó un método para detectar la enfermedad de Parkinson desde espectrogramas de la señal de voz basado en Máquinas de Aprendizaje Extremo.

De los experimentos, se observa que el aumento de muestras, en general, permite mejorar la calidad del entrenamiento, obteniéndose mejores valores de performance. En particular, el experimento que considera espectrogramas en color de fragmentos de sonidos (experimento 3) resulta ser el que tiene mejor resultado. Además de lo anterior, se concluye que el clasificador basado en Máquinas de Aprendizaje Extremo alcanza un nivel de precisión similar al de las Redes Neuronales Convolucionales pero con un tiempo de entrenamiento sustancialmente más reducido.

5.2. Trabajos Futuros

Como es sabido, las técnicas de DL necesitan gran cantidad de muestras etiquetadas para el entrenamiento del modelo. Aunque se han estudiado técnicas de aumentación de datos, sería de utilidad contar con una mayor cantidad de muestras en la base almacenada en el repositorio de la UNLaM. Por tal motivo, es nuestro interés aumentar el número de muestras de voces capturadas en ambiente controlado.

Si bien las grabaciones capturadas en ambientes controlados han producido resultados aceptables para la clasificación de EP y controles sanos, tienen el inconveniente de requerir el traslado del paciente y el uso de equipamiento profesional a cargo de personal capacitado. Como los síntomas del EP se van intensificando progresivamente llegando hasta la invalidez, el traslado se vuelve un problema no menor. No todos los enfermos tienen la posibilidad de contar con un familiar o asistente que pueda trasladarlos a la sede donde se realizan las terapias. Más aún, muchos de ellos necesitan un medio de transporte acorde a sus circunstancias particulares. El uso de dispositivos móviles para capturar los audios de los pacientes representa una oportunidad pues el procedimiento es rá-

pido, no invasivo y no requiere su traslado, sino que podría seguir instrucciones y enviar la grabación.

En un futuro cercano, se propone la creación de una base de datos de grabaciones utilizando dispositivos móviles en ambiente natural, lo que permitirá obtener un mayor número de muestras. Se llevarán a cabo estudios sobre esta base de datos con el objetivo de validar el uso de dispositivos móviles para la adquisición de sonidos en ambientes naturales, y analizar su rendimiento en la clasificación de la Enfermedad de Parkinson.

En un futuro más mediato, se propone estudiar otros tipos de espectrogramas para comparar su efectividad en el proceso de clasificación.

5.3. Publicaciones

A continuación se detallan los trabajos realizados a partir de los resultados obtenidos en esta tesis:

- Mayo 2020 - ”Predicción de la enfermedad de Parkinson utilizando redes neuronales convolucionales”. - Guatelli, R., Aubin, V. I., Pérez, S. N.. In XXII Workshop de Investigadores en Ciencias de la Computación (WICC 2020, El Calafate, Santa Cruz).
- Octubre 2021 - ”Detección de Parkinson mediante espectrogramas en color y redes neuronales convolucionales”. - Guatelli, R., Aubin, V. I., Mora, M., Naranjo-Torres, J., Sinopoli, A. In II Simposio Argentino de Imágenes y Visión (SAIV 2021)-JAIIO 50 (Modalidad virtual).
- Enero 2022 - ”Classification of Parkinson’s disease patients based on spectrogram using local binary pattern descriptors”. - E Gelvez-Almeida, A Vásquez-Coronel, R Guatelli, V Aubin and M Mora. In Journal of Physics: Conference Series. IOP Publishing.

- Noviembre 2022 - ”Análisis comparativo entre CNN y Modelos Logísticos para detección de la Enfermedad de Parkinson utilizando la voz” - Renata S. Guatelli, Monica Giuliano, Verónica Aubin, Luis Fernández, María Laura Pepe, Silvia N. Perez - In 10º Congreso Nacional de Ingeniería Informática / Sistemas de Información (CoNaIISI 2022).
- Julio 2023 - ”Detection of Parkinson’s disease based on spectrograms of voice recordings and Extreme Learning Machine random weight neural networks”. Autores: Renata Guatelli, Verónica Aubin, Marco Mora, Jose Naranjo-Torres, Antonia Mora-Olivari - Journal: Engineering Applications of Artificial Intelligence, 125, 106700.

.

Página intencionalmente en blanco

Referencias

[1] Sustancia negra y el mal de parkinson. URL: https://medlineplus.gov/spanish/ency/esp_imagepages/19515.htm, 2022. (Consultado en marzo de 2023).

[2] P. Deepan and L. R. Sudha. *Deep Learning Algorithm and Its Applications to IoT and Computer Vision*, pages 223–244. Springer Singapore, Singapore, 2021.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[4] Vgg16 model architecture. URL: https://www.kaggle.com/code/yasserhessein/gender-classification-using-vgg16-cnn, 2014. (Consultado en marzo de 2023).

[5] Wan-Jung Chang, Liang-Bi Chen, Chia-Hao Hsu, Cheng-Pei Lin, and Tzu-Chin Yang. A deep learning-based intelligent medicine recognition system for chronic patients. *IEEE Access*, 7:44441–44458, 2019.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and<0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[8] Inc. MathWorks. Deep learning toolbox™ - matlab. https://www.mathworks.com/products/deep-learning.html, 2020.

[9] Jeri A Logemann, Hilda B Fisher, Benjamin Boshes, and E Richard Blonsky. Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients. *Journal of Speech and hearing Disorders*, 43(1):47–57, 1978.

[10] Joseph Jankovic. Parkinson's disease: clinical features and diagnosis. *Journal of neurology, neurosurgery & psychiatry*, 79(4):368–376, 2008.

[11] Martin Köllensperger, Felix Geser, Klaus Seppi, Michaela Stampfer-Kountchev, Martin Sawires, Christoph Scherfler, Sylvia Boesch, Joerg Mueller, Vasiliki Koukouni, Niall Quinn, et al. Red flags for multiple system atrophy. *Movement disorders: official journal of the Movement Disorder Society*, 23(8):1093–1099, 2008.

[12] Francisco Javier Carod-Artal, Hudson Mourão Mesquita, Sofia Ziomkowski, and Pablo Martinez-Martin. Burden and health-related quality of life among caregivers of brazilian parkinson's disease patients. *Parkinsonism & related disorders*, 19(11):943–948, 2013.

[13] AH Schapira, Y Agid, P Barone, P Jenner, MR Lemke, W Poewe, O Rascol, H Reichmann, and E Tolosa. Perspectives on recent advances in the understanding and treatment of parkinson's disease. *European journal of neurology*, 16(10):1090–1099, 2009.

[14] Athanasios Tsanas, Max A Little, Patrick E McSharry, Jennifer Spielman, and Lorraine O Ramig. Novel speech signal processing algorithms for high-accuracy classification of parkinson's disease. *IEEE transactions on biomedical engineering*, 59(5):1264–1271, 2012.

[15] Siddharth Arora, Ladan Baghai-Ravary, and Athanasios Tsanas. Developing a large scale population screening tool for the assessment of parkinson's disease using telephone-quality voice. *The Journal of the Acoustical Society of America*, 145(5):2871–2884, 2019.

[16] Laetitia Jeancolas, Habib Benali, Badr-Eddine Benkelfat, Graziella Mangone, Jean-Christophe Corvol, Marie Vidailhet, Stéphane Lehericy, and Dijana Petrovska-Delacrétaz. Automatic detection of early stages of parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients. In *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–6. IEEE, 2017.

[17] Rhonda J. Holmes, Jennifer M. Oates, Debbie J. Phyland, and Andrew J. Hughes. Voice characteristics in the progression of parkinson's disease. *International Journal of Language & Communication Disorders*, 35(3):407–418, 2000.

[18] Ervin Sejdic, Igor Djurovic, and LJubisa Stankovic. Quantitative performance analysis of scalogram as instantaneous frequency estimator. *IEEE Transactions on Signal Processing*, 56(8):3837–3845, 2008.

[19] Brian ED Kingsbury, Nelson Morgan, and Steven Greenberg. Robust speech recognition using the modulation spectrogram. *Speech communication*, 25(1-3):117–132, 1998.

[20] Steven Greenberg and Brian Kingsbury. The modulation spectrogram: in pursuit of an invariant representation of speech. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3:1647–1650 vol.3, 1997.

[21] Mahmood Saleh Alzubaidi, Uzair Shah, Haider Dhia Zubaydi, Khalid Dolaat, Alaa A. Abd-Alrazaq, Arfan Ahmed, and Mowafa Househ. The role of neural network for the detection of parkinson's disease: A scoping review. *Healthcare*, 9(6):740, jun 2021.

[22] Máté Hireš, Matej Gazda, Peter Drotár, Nemuel Daniel Pah, Mohammod Abdul Motin, and Dinesh Kant Kumar. Convolutional neural network ensemble for parkinson's disease detection from voice recordings. *Computers in biology and medicine*, 141:105021, 2022.

[23] Manisha Jindal and Yogesh Tripathi. Parkinson's disease detection using convolutional neural networks. *Eur. J. Mol. Clin. Med*, 7(6):1298–1307, 2020.

[24] Shivangi, Anubhav Johri, and Ashish Tripathi. Parkinson disease detection using deep neural networks. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*. IEEE, aug 2019.

[25] Jihun Kim, Jonghong Kim, Gil-Jin Jang, and Minho Lee. Fast learning method for convolutional neural networks using extreme learning machine and its application to lane detection. *Neural Networks*, 87:109–121, mar 2017.

[26] Youngwoo Yoo and Se-Young Oh. Fast training of convolutional neural network classifiers through extreme learning machines. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, jul 2016.

[27] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[28] Gao Huang, Shiji Song, Jatinder N. D. Gupta, and Cheng Wu. Semi-supervised and unsupervised extreme learning machines. *IEEE Transactions on Cybernetics*, 44(12):2405–2417, 2014.

[29] Jichao Chen, Yijie Zeng, Yue Li, and Guang-Bin Huang. Unsupervised feature selection based extreme learning machine for clustering. *Neurocomputing*, 386:198–207, 2020.

[30] Eliakim H Moore. On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.*, 26:394–395, 1920.

[31] R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, jul 1955.

[32] Xin Bi, Xiangguo Zhao, Hong Huang, Deyang Chen, and Yuliang Ma. Functional brain network classification for alzheimer's disease detection

with deep features and extreme learning machine. *Cognitive Computation*, 12:513–527, 2020.

[33] Hui-Ling Chen, Gang Wang, Chao Ma, Zhen-Nao Cai, Wen-Bin Liu, and Su-Jing Wang. An efficient hybrid kernel extreme learning machine approach for early diagnosis of parkinson's disease. *Neurocomputing*, 184:131–144, 2016.

[34] Qi Liu, Xiaoguang Zhao, Zeng guang Hou, and Hongguang Liu. Epileptic seizure detection based on the kernel extreme learning machine. *Technology and health care : official journal of the European Society for Engineering and Medicine*, 25 S1:399–409, 2017.

[35] Vivek Lahoura, Harpreet Singh, Ashutosh Aggarwal, Bhisham Sharma, Mazin Abed Mohammed, Robertas Damaševičius, Seifedine Kadry, and Korhan Cengiz. Cloud computing-based framework for breast cancer diagnosis using extreme learning machine. *Diagnostics*, 11(2):241, 2021.

[36] Farhat Afza, Muhammad Sharif, Muhammad Attique Khan, Usman Tariq, Hwan-Seung Yong, and Jaehyuk Cha. Multiclass skin lesion classification using hybrid deep features selection and extreme learning machine. *Sensors (Basel, Switzerland)*, 22, 2022.

[37] Vincent P Calabrese. Projected number of people with parkinson disease in the most populous nations, 2005 through 2030. *Neurology*, 69(2):223–224, 2007.

[38] James Parkinson. An essay on the shaking palsy. *The Journal of neuropsychiatry and clinical neurosciences*, 14(2):223–236, 2002.

[39] James R Williamson, Thomas F Quatieri, Brian S Helfer, Joseph Perricone, Satrajit S Ghosh, Gregory Ciccarelli, and Daryush D Mehta. Segment-dependent dynamics in predicting parkinson's disease. In *Sixteenth annual conference of the international speech communication association*, 2015.

[40] Janice M Beitz. Parkinson's disease: a review. *Frontiers in Bioscience-Scholar*, 6(1):65–74, 2014.

[41] Ledia F Hernández, Peter Redgrave, and José A Obeso. Habitual behavior and dopamine cell vulnerability in parkinson disease, 2015.

[42] O Castillero Mimenza. Ganglios basales: anatomía y funciones. *Psicología y Mente*, 2016. (Consultado en marzo de 2023).

[43] Panchanan Maiti, Jayeeta Manna, and Gary L Dunbar. Current understanding of the molecular mechanisms in parkinson's disease: targets for potential treatments. *Translational neurodegeneration*, 6:1–35, 2017.

[44] Tanya Simuni, Chelsea Caspell-Garcia, Christopher S Coffey, Daniel Weintraub, Brit Mollenhauer, Shirley Lasch, Caroline M Tanner, Danna Jennings, Karl Kieburtz, Lana M Chahine, et al. Baseline prevalence and longitudinal evolution of non-motor symptoms in early parkinson's disease: the ppmi cohort. *Journal of Neurology, Neurosurgery & Psychiatry*, 89(1):78–88, 2018.

[45] Werner Poewe. Non-motor symptoms in parkinson's disease. *European journal of neurology*, 15:14–20, 2008.

[46] K Berganzo, B Tijero, A González-Eizaguirre, J Somme, E Lezcano, I Gabilondo, M Fernandez, JJ Zarranz, and JC Gómez-Esteban. Síntomas no motores y motores en la enfermedad de parkinson y su relación con la calidad de vida y los distintos subgrupos clínicos. *Neurología*, 31(9):585–591, 2016.

[47] Carmen Gil and Ana Martínez. *El parkinson*. LOS LIBROS DE LA CATARATA, 2016.

[48] Federico Micheli. *Tratado de neurología clínica*. Ed. Médica Panamericana, 2002.

[49] Anette Schrag, Yoav Ben-Shlomo, Richard Brown, C David Marsden, and Niall Quinn. Young-onset parkinson's disease revisited—clinical fea-

tures, natural history, and mortality. *Movement disorders: official journal of the Movement Disorder Society*, 13(6):885–894, 1998.

[50] Günther Deuschl, Ettore Beghi, Franz Fazekas, Timea Varga, Kalliopi A Christoforidi, Eveline Sipido, Claudio L Bassetti, Theo Vos, and Valery L Feigin. The burden of neurological diseases in europe: an analysis for the global burden of disease study 2017. *The Lancet Public Health*, 5(10):e551–e567, 2020.

[51] Tamara Pringsheim, Nathalie Jette, Alexandra Frolkis, and Thomas DL Steeves. The prevalence of parkinson's disease: a systematic review and meta-analysis. *Movement disorders*, 29(13):1583–1590, 2014.

[52] Lauren Hirsch, Nathalie Jette, Alexandra Frolkis, Thomas Steeves, and Tamara Pringsheim. The incidence of parkinson's disease: a systematic review and meta-analysis. *Neuroepidemiology*, 46(4):292–300, 2016.

[53] Parra Medina Luis Enrique, Arankowsky Sandoval Gloria, Salazar Ceballos Jorge Efraín, and Góngora Alfaro José Luis. Latencia diagnóstica en la enfermedad de parkinson y su relación con los síntomas prodrómicos motores y no motores. *Eneurobiología*, 10(25):1, 2019.

[54] Mayela Rodríguez-Violante and Amin Cervantes-Arriaga. La escala unificada de la enfermedad de parkinson modificada por la sociedad de trastornos del movimiento (mds-updrs): aplicación clínica e investigación. *Archivos de Neurociencias*, 19(3):157–163, 2014.

[55] Fátima Goulart and Luciana Xavier Pereira. Uso de escalas para avaliação da doença de parkinson em fisioterapia. *Fisioterapia e pesquisa*, 11(1):49–56, 2005.

[56] Margaret M Hoehn, Melvin D Yahr, et al. Parkinsonism: onset, progression, and mortality. *Neurology*, 50(2):318–318, 1998.

[57] Douglas J Gelb, Eugene Oliver, and Sid Gilman. Diagnostic criteria for parkinson disease. *Archives of neurology*, 56(1):33–39, 1999.

[58] John G Nutt and G Frederick Wooten. Diagnosis and initial management of parkinson's disease. *New England Journal of Medicine*, 353(10):1021–1027, 2005.

[59] Ronald B Postuma, Daniela Berg, Matthew Stern, Werner Poewe, C Warren Olanow, Wolfgang Oertel, José Obeso, Kenneth Marek, Irene Litvan, Anthony E Lang, et al. Mds clinical diagnostic criteria for parkinson's disease. *Movement disorders*, 30(12):1591–1601, 2015.

[60] Andrew J Hughes, Yoav Ben-Shlomo, Susan E Daniel, and Andrew J Lees. What features improve the accuracy of clinical diagnosis in parkinson's disease: a clinicopathologic study. *Neurology*, 42(6):1142–1142, 1992.

[61] David Sulzer. Multiple hit hypotheses for dopamine neuron loss in parkinson's disease. *Trends in neurosciences*, 30(5):244–250, 2007.

[62] Georg Becker, Antje Müller, Stefan Braune, Thomas Büttner, Reiner Benecke, Wolfgang Greulich, Wolfgang Klein, Günter Mark, Jürgen Rieke, and Reiner Thümler. Early diagnosis of parkinson's disease. *Journal of neurology*, 249(3):iii40–iii48, 2002.

[63] Giovanni Rizzo, Massimiliano Copetti, Simona Arcuti, Davide Martino, Andrea Fontana, and Giancarlo Logroscino. Accuracy of clinical diagnosis of parkinson disease: a systematic review and meta-analysis. *Neurology*, 86(6):566–576, 2016.

[64] O Tucha, L Mecklinger, J Thome, A Reiter, GL Alders, H Sartor, M Naumann, and KW Lange. Kinematic analysis of dopaminergic effects on skilled handwriting movements in parkinson's disease. *Journal of neural transmission*, 113:609–623, 2006.

[65] Evelien Nackaerts, Griet Vervoort, Elke Heremans, Bouwien CM Smits-Engelsman, Stephan P Swinnen, and Alice Nieuwboer. Relearning of writing skills in parkinson's disease: a literature review on influential

factors and optimal strategies. *Neuroscience & Biobehavioral Reviews*, 37(3):349–357, 2013.

[66] Peter Drotár, Jiří Mekyska, Irena Rektorová, Lucia Masarová, Zdeněk Smékal, and Marcos Faundez-Zanuy. Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. *Artificial intelligence in Medicine*, 67:39–46, 2016.

[67] Hans-Leo Teulings, José L Contreras-Vidal, George E Stelmach, and Charles H Adler. Parkinsonism reduces coordination of fingers, wrist, and arm in fine motor control. *Experimental neurology*, 146(1):159–170, 1997.

[68] Iqra Kamran, Saeeda Naz, Imran Razzak, and Muhammad Imran. Handwriting dynamics assessment using deep neural network for early identification of parkinson's disease. *Future Generation Computer Systems*, 117:234–244, 2021.

[69] C Kotsavasiloglou, Nikolaos Kostikis, Dimitrios Hristu-Varsakelis, and Marianthi Arnaoutoglou. Machine learning-based classification of simple drawing movements in parkinson's disease. *Biomedical Signal Processing and Control*, 31:174–180, 2017.

[70] Clayton R Pereira, Silke AT Weber, Christian Hook, Gustavo H Rosa, and Joao P Papa. Deep learning-aided parkinson's disease diagnosis from handwritten dynamics. In *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–346. Ieee, 2016.

[71] Jun-Xiu Yang and Lei Chen. Economic burden analysis of parkinson's disease patients in china. *Parkinson's Disease*, 2017, 2017.

[72] Tuan D Pham. Pattern analysis of computer keystroke time series in healthy control and early-stage parkinson's disease subjects using fuzzy recurrence and scalable recurrence network features. *Journal of Neuroscience Methods*, 307:194–202, 2018.

[73] Jingjing Xu and Minming Zhang. Use of magnetic resonance imaging and artificial intelligence in studies of diagnosis of parkinson's disease. *ACS chemical neuroscience*, 10(6):2658–2667, 2019.

[74] Bright Chukwunwike Uzuegbunam, Damiano Librizzi, and Behrooz Hooshyar Yousefi. Pet radiopharmaceuticals for alzheimer's disease and parkinson's disease diagnosis, the current and future landscape. *Molecules*, 25(4):977, 2020.

[75] Farhan Mohammed, Xiangjian He, and Yiguang Lin. Retracted: An easy-to-use deep-learning model for highly accurate diagnosis of parkinson's disease using spect images, 2021.

[76] Peter F MacNeilage. *The origin of speech*. Oxford University Press, 2010.

[77] Ignacio Cobeta, Faustino Núñez, and Secundino Fernández. *Patología de la voz*. Marge books, 2013.

[78] Robert T Sataloff, Yolanda D Heman-Ackah, and Mary J Hawkshaw. Clinical anatomy and physiology of the voice. *Otolaryngologic clinics of north America*, 40(5):909–929, 2007.

[79] J Gamboa, FJ Jiménez Jiménez, MA Mate, and I Cobeta. Alteraciones de la voz causadas por enfermedades neurológicas. *Rev. neurol.(Ed. impr.)*, pages 153–168, 2001.

[80] Ingo R Titze. Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, 85(4):1699–1707, 1989.

[81] Carole T Ferrand. Speech science: An integrated approach to theory and clinical practice. *Ear and Hearing*, 22(6):549, 2001.

[82] Shaheen N Awan and Michael L Frenkel. Improvements in estimating the harmonics-to-noise ratio of the voice. *Journal of Voice*, 8(3):255–262, 1994.

[83] J Peter H Pabon and Reinier Plomp. Automatic phonetogram recording supplemented with acoustical voice-quality parameters. *Journal of Speech, Language, and Hearing Research*, 31(4):710–722, 1988.

[84] Jiri Mekyska, Irena Rektorova, and Zdenek Smekal. Selection of optimal parameters for automatic analysis of speech disorders in parkinson's disease. In *2011 34th International Conference on Telecommunications and Signal Processing (TSP)*, pages 408–412. IEEE, 2011.

[85] Natalia Gabriela Elisei. Análisis acústico de la voz normal y patológica utilizando dos sistemas diferentes: Anagraf y praat. *Interdisciplinaria*, 29(2):271–286, 2012.

[86] Brian T Harel, Michael S Cannizzaro, Henri Cohen, Nicole Reilly, and Peter J Snyder. Acoustic characteristics of parkinsonian speech: a potential biomarker of early disease progression and treatment. *Journal of Neurolinguistics*, 17(6):439–453, 2004.

[87] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[88] Juan Rafael Orozco-Arroyave, Julián David Arias-Londoño, Jesús Francisco Vargas-Bonilla, María Claudia Gonzalez-Rátiva, and Elmar Nöth. New spanish speech corpus database for the analysis of people suffering from parkinson's disease. In *LREC*, pages 342–347, 2014.

[89] Laura Fogg and Alison Talmage. The celebration choir: Establishing community group choral singing for people living with neurological conditions. *Psychomusicology: Music, Mind and Brain*, 21(1-2):264, 2011.

[90] Lorraine O Ramig, Shimon Sapir, Cynthia Fox, and Stefanie Countryman. Changes in vocal loudness following intensive voice treatment (lsvt®) in individuals with parkinson's disease: A comparison with un-

treated patients and normal age-matched controls. *Movement disorders: official journal of the Movement Disorder Society*, 16(1):79–83, 2001.

[91] Secundino Fernández González, D Ruba San Miguel, M Marqués Girbau, and L Sarraqueta. Voz del anciano. *Revista de Medicina de la Universidad de Navarra*, pages 44–48, 2006.

[92] Murray Morrison, Linda Rammage, Hamish Nichol, Bruce Pullan, Phillip May, and Lesley Salkeld. *Tratamiento de los trastornos de la voz*. Masson, 2006.

[93] Jeffrey L Cummings and Donna L Masterman. Depression in patients with parkinson's disease. *International journal of geriatric psychiatry*, 14(9):711–718, 1999.

[94] Nick Miller, Emma Noble, Diana Jones, Liesl Allcock, and David J Burn. How do i sound to me? perceived changes in communication in parkinson's disease. *Clinical rehabilitation*, 22(1):14–22, 2008.

[95] KA Flowers, C Robertson, and MR Sheridan. Some characteristics of word fluency in parkinson's disease. *Journal of Neurolinguistics*, 9(1):33–46, 1995.

[96] Sabine Skodda and Uwe Schlegel. Speech rate and rhythm in parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, 23(7):985–992, 2008.

[97] Hermann Ackermann and Wolfram Ziegler. Articulatory deficits in parkinsonian dysarthria: an acoustic analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, 54(12):1093–1098, 1991.

[98] Adolfo M García, Facundo Carrillo, Juan Rafael Orozco-Arroyave, Natalia Trujillo, Jesús F Vargas Bonilla, Sol Fittipaldi, Federico Adolfi, Elmar Nöth, Mariano Sigman, Diego Fernández Slezak, et al. How language flows when movements don't: an automated analysis of spontaneous discourse in parkinson's disease. *Brain and language*, 162:19–28, 2016.

[99] Nick Miller, Liesl Allcock, Diana Jones, Emma Noble, Anthony J Hildreth, and David J Burn. Prevalence and pattern of perceived intelligibility changes in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 78(11):1188–1190, 2007.

[100] Marc D Pell, Henry S Cheang, and Carol L Leonard. The impact of parkinson's disease on vocal-prosodic communication from the perspective of listeners. *Brain and language*, 97(2):123–134, 2006.

[101] Sally Gallena, Paul J Smith, Thomas Zeffiro, and Christy L Ludlow. Effects of levodopa on laryngeal muscle activity for voice onset and offset in parkinson disease. *Journal of speech, language, and hearing research : JSLHR*, 44 6:1284–99, 2001.

[102] Judith B King, Lorraine Olson Ramig, Jon H Lemke, and Yoshiyuki Horii. Parkinson's disease: longitudinal changes in acoustic parameters of phonation. *NCVS Status Prog Rep*, 4:135–149, 1993.

[103] E Jeffrey Metter and Wayne R Hanson. Clinical and acoustical variability in hypokinetic dysarthria. *Journal of communication disorders*, 19(5):347–366, 1986.

[104] Betul Erdogdu Sakar, M Erdem Isenkul, C Okan Sakar, Ahmet Sertbas, Fikret Gurgen, Sakir Delil, Hulya Apaydin, and Olcay Kursun. Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834, 2013.

[105] Meysam Asgari and Izhak Shafran. Extracting cues from speech for predicting severity of parkinson's disease. In *2010 IEEE international workshop on machine learning for signal processing*, pages 462–467. IEEE, 2010.

[106] C Okan Sakar and Olcay Kursun. Telediagnosis of parkinson's disease using measurements of dysphonia. *Journal of medical systems*, 34:591–599, 2010.

[107] Musaed Alhussein and Ghulam Muhammad. Voice pathology detection using deep learning on mobile healthcare framework. *IEEE Access*, 6:41034–41041, 2018.

[108] Athanasios Tsanas, Max A Little, Patrick E McSharry, and Lorraine O Ramig. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average parkinson's disease symptom severity. *Journal of the royal society interface*, 8(59):842–855, 2011.

[109] Francisco Díaz-Pérez, Evelyn García-Nieto, Antonio Ros, and Rafael Claramunt. Best estimation of spectrum profiles for diagnosing femoral prostheses loosening. *Medical Engineering & Physics*, 36(2):233–238, 2014.

[110] David Montaña, Yolanda Campos-Roca, and Carlos J Pérez. A diadochokinesis-based expert system considering articulatory features of plosive consonants for early detection of parkinson's disease. *Computer methods and programs in biomedicine*, 154:89–97, 2018.

[111] Félix Javier Jiménez-Jiménez, Javier Gamboa, Alberto Nieto, Josana Guerrero, Miguel Orti-Pareja, Jose Antonio Molina, Esteban García-Albea, and Ignacio Cobeta. Acoustic voice analysis in untreated patients with parkinson's disease. *Parkinsonism & Related Disorders*, 3(2):111–116, 1997.

[112] A.J. Flint, S.E. Black, I. Campbell-Taylor, G.F. Gailey, and C. Levinton. Acoustic analysis in the differentiation of parkinson's disease and major depression. *Journal of Psycholinguistic Research*, 21, 383-399., 1992.

[113] Ingo Hertrich and Hermann Ackermann. Gender-specific vocal dysfunctions in parkinson's disease: electroglottographic and acoustic analyses. *Annals of Otology, Rhinology & Laryngology*, 104(3):197–202, 1995.

[114] OR Aguilera Pacheco, DI Escobedo Beceiro, F Sanabria Macias, and I Nuñez Lahera. Alteración de parámetros acústicos de la voz y el ha-

bla en la enfermedad de parkinson. In *XIV Simposio Internacional de Comunicación Social. Comunicación Social: Retos y Perspectivas*, volume 2, 2015.

[115] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[116] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678, 2014.

[117] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[118] WJ Barry and M Pützer. Saarbrucken voice database. *Institute of Phonetics*, 2007.

[119] Juan Camilo Vásquez-Correa, Tomas Arias-Vergara, Cristian D Rios-Urrego, Maria Schuster, Jan Rusz, Juan Rafael Orozco-Arroyave, and Elmar Nöth. Convolutional neural networks and a transfer learning strategy to classify parkinson's disease from speech in three different languages. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 24th Iberoamerican Congress, CIARP 2019, Havana, Cuba, October 28-31, 2019, Proceedings 24*, pages 697–706. Springer, 2019.

[120] Marek Wodzinski, Andrzej Skalski, Daria Hemmerling, Juan Rafael Orozco-Arroyave, and Elmar Nöth. Deep learning approach to parkinson's disease detection using voice recordings and convolutional neural network dedicated to image classification. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 717–720. IEEE, 2019.

[121] Laiba Zahid, Muazzam Maqsood, Mehr Yahya Durrani, Maheen Bakhtyar, Junaid Baber, Habibullah Jamal, Irfan Mehmood, and Oh-Young Song. A spectrogram-based deep feature assisted computer-aided diagnostic system for parkinson's disease. *IEEE Access*, 8:35482–35495, 2020.

[122] Nam Trinh and O'Brien Darragh. Pathological speech classification using a convolutional neural network. ., 2019.

[123] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[124] DAR Venegas. Dataset of vowels. URL: `https://www.kaggle.com/datasets/darubiano57/dataset-of-vowels`, 2018. Accedido el 9 de marzo de 2023.

[125] Anubhav Johri, Ashish Tripathi, et al. Parkinson disease detection using deep neural networks. In *2019 Twelfth international conference on contemporary computing (IC3)*, pages 1–4. IEEE, 2019.

[126] David E. Rumelhart and James L. McClelland. Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations. In *microstructureofcognition*, 1986.

[127] James L McClelland, David E Rumelhart, PDP Research Group, et al. *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*, volume 2. MIT press, 1987.

[128] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. *Advances in neural information processing systems*, 25, 2012.

[129] Koray Kavukcuoglu, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michaël Mathieu, Yann Cun, et al. Learning convolutional feature hierarchies for visual recognition. *Advances in neural information processing systems*, 23, 2010.

[130] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[131] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and trends® in signal processing*, 7(3–4):197–387, 2014.

[132] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer, 2018.

[133] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[134] Le Zhang and Ponnuthurai N Suganthan. A survey of randomized algorithms for training neural networks. *Information Sciences*, 364:146–155, 2016.

[135] Ponnuthurai N Suganthan and Rakesh Katuwal. On the origins of randomization-based feedforward neural networks. *Applied Soft Computing*, 105:107239, 2021.

[136] David S Broomhead and David Lowe. Radial basis functions, multivariable functional interpolation and adaptive networks. Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

[137] David Lowe. Adaptive radial basis function nonlinearities, and the problem of generalisation. In *1989 First IEE International Conference on Artificial Neural Networks,(Conf. Publ. No. 313)*, pages 171–175. IET, 1989.

[138] Wouter F Schmidt, Martin A Kraaijveld, Robert PW Duin, et al. Feed forward neural networks with random weights. In *International conference on pattern recognition*, pages 1–1. IEEE Computer Society Press, 1992.

[139] Yoh-Han Pao, Gwang-Hoon Park, and Dejan J Sobajic. Learning and generalization characteristics of the random vector functional-link net. *Neurocomputing*, 6(2):163–180, 1994.

[140] Hugues Berry and Mathias Quoy. Structure and dynamics of random recurrent neural networks. *Adaptive Behavior*, 14(2):129–137, 2006.

[141] Ashwani Kumar Malik, Ruobin Gao, MA Ganaie, Muhammad Tanveer, and Ponnuthurai N Suganthan. Random vector functional link network: recent developments, applications, and future directions. *arXiv preprint arXiv:2203.11316*, 2022.

[142] Guang-Bin Huang, Lei Chen, Chee Kheong Siew, et al. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. Neural Networks*, 17(4):879–892, 2006.

[143] Guang-Bin Huang and Lei Chen. Convex incremental extreme learning machine. *Neurocomputing*, 70(16-18):3056–3062, 2007.

[144] Guang-Bin Huang and Lei Chen. Enhanced random search based incremental extreme learning machine. *Neurocomputing*, 71(16-18):3460–3468, 2008.

[145] Guang-Bin Huang, Hongming Zhou, Xiaojian Ding, and Rui Zhang. Extreme learning machine for regression and multiclass classification. *IEEE*

Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(2):513–529, 2011.

[146] Peter Adeniyi Alaba, Segun Isaiah Popoola, Lanre Olatomiwa, Mathew Boladele Akanle, Olayinka S Ohunakin, Emmanuel Adetiba, Opeoluwa David Alex, Aderemi AA Atayero, and Wan Mohd Ashri Wan Daud. Towards a more efficient and cost-sensitive extreme learning machine: A state-of-the-art review of recent trend. *Neurocomputing*, 350:70–90, 2019.

[147] Wan-Yu Deng, Qing-Hua Zheng, Shiguo Lian, Lin Chen, and Xin Wang. Ordinal extreme learning machine. *Neurocomputing*, 74(1-3):447–456, 2010.

[148] C Radhakrishna Rao and Sujit Kumar Mitra. Further contributions to the theory of generalized inverse of matrices and its applications. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 289–300, 1971.

[149] Monica Giuliano, Silvia Noemi Perez, Maldonado Maldonado, Pablo Bondar, Daniela Linari, Dario Adamec Adamec, María Inés Debas, Carlos Morales Morales, Leticia de León, Aldo Yaco Yaco, Joice Birelli Birelli, Macarena Martínez Ribaya, María Lis Lacaze, and Jorge A. Gurlekian. Construction of a parkinson´s voice database. In *International Conference on Industrial Engineering and Operations Management*. IEOM Society International, 2021.

[150] Christopher G Goetz, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Glenn T Stebbins, Matthew B Stern, Barbara C Tilley, Richard Dodel, Bruno Dubois, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): process, format, and clinimetric testing plan. *Movement disorders*, 22(1):41–47, 2007.

[151] Yiting Xie and David Richmond. Pre-training on grayscale imagenet improves medical image classification. In *Proceedings of the European*

conference on computer vision (ECCV) workshops, pages 0–0, 2018.

[152] Sagar Sharma, Simone Sharma, and Anidhya Athaiya. Activation functions in neural networks. *Towards Data Sci*, 6(12):310–316, 2017.

[153] Quoc Cuong Ngo, Mohammod Abdul Motin, Nemuel Daniel Pah, Peter Drotár, Peter Kempster, and Dinesh Kumar. Computerized analysis of speech and voice for parkinson's disease: A systematic review. *Computer Methods and Programs in Biomedicine*, page 107133, 2022.

[154] Monica Giuliano, Luis Fernandez, and Silvia Pérez. Selección de medidas de disfonía para la identificación de enfermos de parkinson. In *2020 IEEE Congreso Bienal de Argentina (ARGENCON)*, pages 1–8. IEEE, 2020.

[155] Athanasios Tsanas. *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning*. PhD thesis, Oxford University, UK, 2012.

[156] A Tsanas. Automatic objective biomarkers of neurodegenerative disorders using nonlinear speech signal processing tools. In *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, pages 37–40, 2013.

[157] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.