UNIVERSIDAD
DE ANTIOQUIA

1 8 0 3

Aprendizaje por transferencia en redes neuronales convolucionales para el diagnóstico y monitoreo de la enfermedad de Parkinson usando señales de voz en tres idiomas diferentes.

Cristian David Rios Urrego.

Universidad de Antioquia
Facultad de Ingeniería
Departamento de Ingeniería Electrónica y Telecomunicaciones
Medellín, Colombia
2019

Aprendizaje por transferencia en redes neuronales convolucionales para el diagnóstico y monitoreo de la enfermedad de Parkinson usando señales de voz en tres idiomas diferentes.

Cristian David Rios Urrego
Estudiante de pregrado en Ingeniería Electrónica

Trabajo de grado presentado para optar por el título de:
Ingeniero Electrónico

Asesor:
PhD. Juan Rafael Orozco Arroyave

Co-Asesor:
MSc. Juan Camilo Vásquez Correa

Linea de investigación:
Procesamiento digital de señales y análisis de patrones
Grupo de Investigación en Telecomunicaciones Aplicadas
GITA

Universidad de Antioquia
Facultad de Ingeniería
Departamento de Ingeniería Electrónica y Telecomunicaciones
Medellín, Colombia
2019

Agradecimientos

Primero quiero agradecerle a mi madre Beatriz Urrego, a mi padre Hernán Rios y a mi hermano Camilo Rios, quienes me han apoyado en este proceso y han sido mis pilares para formarme como una persona íntegra. También agradecerle a gran parte de mi familia, los cuales han dejado su grano de arena y han confiado en mi para la culminación de mi carrera.

Agradezco a mis compañeros de pregrado Felipe López y Daniel Escobar, los cuales me han acompañado en gran parte de mi vida académica. También a mis compañeros y amigos del grupo de investigación GITA: Paula Pérez, Surley Berrio, Felipe Gómez, Felipe Parra, Tomás Arias y Nicanor García por guiarme y acompañarme en este proceso académico.

Por último, quiero agradecer a mi tutor el profesor Juan Rafael Orozco y a mi co-asesor Juan Camilo Vásquez, por el aprendizaje que pude obtener a través de sus experiencias, por la disposición y por la paciencia en este largo proceso de investigación.

Índice

Índice de figuras

Índice de tablas

Resumen

La enfermedad de Parkinson es un desorden neurodegenerativo del sistema nervioso caracterizado por rigidez, bradicinesia y pérdida de los reflejos posturales, afectando drásticamente la calidad de vida de la persona que la padece. Las deficiencias del habla son comúnmente uno de los síntomas tempranos de la enfermedad, por lo que puede ser un buen bio-marcador para el apoyo diagnóstico y el monitoreo de la enfermedad. Este trabajo propone un estudio a partir del aprendizaje profundo, exactamente en la técnica de aprendizaje por transferencia con el fin de mejorar la eficacia de los sistemas para el apoyo diagnóstico de la enfermedad de Parkinson en tres idiomas diferentes: Español, Alemán y Checo. Inicialmente se extrajeron las transiciones de las señales de voz, con el fin de modelar las anomalías que presentan los pacientes para comenzar y/o detener la vibración de los pliegues vocales. Luego estas transiciones se llevaron a una representación tiempo-frecuencia utilizando la Transformada de Fourier de Tiempo Corto, formando espectrogramas muestreados en la escala de Mel, los cuales se usan para el entrenamiento y la validación de redes neuronales convolucionales. Posteriormente con el fin de comparar y comprobar si el aprendizaje por transferencia entre idiomas puede mejorar el apoyo diagnóstico de la enfermedad de Parkinson, se realizaron 4 experimentos diferentes: **(i)** Clasificación con máquinas de soporte vectorial a partir de características articulatorias clásicas. **(ii)** Entrenamiento y evaluación de redes neuronales convolucionales con datos monolingües. **(iii)** Entrenamiento y evaluación de redes neuronales convolucionales implementando aprendizaje por transferencia entre idiomas. **(iv)** Clasificación multiclase con redes neuronales convolucionales para evaluar el estado neurológico de los pacientes.

Los resultados indican que es posible mejorar los modelos monolingües a partir de otros idiomas usando el método de aprendizaje por transferencia, por lo tanto el entrenamiento de los modelos no debe comenzar con parámetros aleatorios como se realiza comúnmente, si no con un modelo base entrenado en un idioma diferente, aunque es necesario que este modelo sea lo suficientemente robusto para realizar una correcta transferencia de conocimiento y poder incrementar el desempeño del sistema transferido.

1. Introducción

1.1. Contexto

La enfermedad de Parkinson (EP) es un trastorno neurológico caracterizado por la pérdida progresiva de neuronas dopaminérgicas en la sustancia nigra del cerebro medio, produciendo déficits motores y no motores en los pacientes, como bradicinesia, rigidez, inestabilidad postural, temblor en reposo, efectos negativos en el sueño, cambios de humor, etc. [1], [2] La escala estándar para evaluar el estado neurológico de los pacientes es la MDS-UPDRS (del inglés Movement Disorder Society-Unified Parkinson's Disease Rating Scale) [3]. La tercera sección (MDS-UPDRS-III) de la escala mide las deficiencias del paciente evaluando aspectos motores de la EP. Esta escala solo contiene 2 ítems que consideran el habla, a pesar de que los desórdenes en la voz son una manifestación temprana y prominente que puede contribuir principalmente al diagnóstico de la EP [4]-[6] . Aproximadamente el 90 % de los pacientes con EP desarrollan diferentes discapacidades del habla, incluida la reducción del volumen, habla monótona, voz entrecortada, articulación imprecisa, etc. Todos estos síntomas pueden agruparse y se denominan disartria hipocinética [2]. La disartria hipocinética comúnmente está relacionada con la reducción de la velocidad de los movimientos articulatorios, es decir, el desplazamiento de los labios, la lengua y la mandíbula suelen ser más, lentos afectando el habla inteligible de las personas y reduciendo drásticamente la habilidad de comunicación de los pacientes con EP [7]. Gran cantidad de los síntomas mencionados anteriormente son controlados a partir de medicamentos, los cuales reducen considerablemente el movimiento de los pacientes, pero no existe evidencia alguna que garantice un efecto positivo de esta medicación en los problemas de la voz [2], solo con terapias realizadas por fonoaudiólogos es posible combatir las habilidades orales de los pacientes.

Una posible solución para esta y demás problemáticas son las herramientas computacionales, las cuales tiene un gran potencial para el apoyo diagnóstico, el monitoreo de la enfermedad y la terapia de pacientes, ya que utiliza métodos no invasivos como lo son las señales de voz y no existe la necesidad de desplazarse donde un especialista. Gran parte de este desarrollo lo ha generado métodos de aprendizaje profundo, los cuales han permitido crear modelos lo suficientemente robustos para muchas aplicaciones. En el momento existen buenos modelos para la clasificación de pacientes con EP y controles sanos independientes de idiomas como los creados en [8]; sin embargo, aún no se conoce qué tanto podría complementar el entrenamiento de estos modelos con datos de un idioma diferente. En este trabajo se implementó un método de deep learning llamado aprendizaje por transferencia, el

cual nos permite mejorar la eficacia de los modelos monolingües, teniendo como base un sistema entrenado en un idioma diferente.

1.2. Estado del arte

Diferentes estudios en la literatura han analizado el déficit articulatorio en pacientes con EP. En [9] el autor propuso la clasificación automática de pacientes con EP y personas sanas a partir de los coeficientes cepstrales en las frecuencias de Mel (MFCC) y la energía en las bandas de Bark. Esta caracterización se hizo en las transiciones entre segmentos sonoros a no sonoros y viceversa, con el fin de modelar las dificultades que presentan los pacientes para comenzar y detener la vibración de las cuerdas vocales. Para este análisis se consideraron grabaciones de voz de textos leídos y monólogos en tres diferentes idiomas (Español, Alemán, Checo). Los autores reportaron resultados del 91 % al 98 % dependiendo del idioma. Sin embargo los resultados de este trabajo son optimistas pues los híper-parámetros del algoritmo de clasificación utilizado fueron optimizados a partir del conjunto de test. En [10] se propuso diseñar un sistema experto para la detección temprana de la EP, realizando un análisis articulatorio a partir de tareas diadochokinéticas (DDKs) como la repetición rápida de silabas como /pa-ta-ka/, calculando características temporales y espectrales extraídas en los segmentos de tiempo de inicio de voz (del inglés Voice Onset Time, VOT). Los autores crearon una base de datos de 27 pacientes con EP, los cuales se encontraban entre 1 y máximo 2.5 en la escala de Hoehn y Yahr (H&Y) [11], es decir, un estado leve de la enfermedad, y 27 controles sanos. En este trabajo sobresalió el análisis de la consonante plosiva /k/ en comparación con las consonantes /p/ y /t/, con una eficiencia de 92.2 % usando máquinas de soporte vectorial. En [8] se realizó un modelamiento del déficit articulatorio en pacientes con EP a partir de redes neuronales convolucionales (del inglés convolutional neural network, CNN). Inicialmente se detectaron las transiciones introducidas en [9] y los mismos idiomas fueron utilizados. Luego se realizó una representación en tiempo-frecuencia a partir de la transformada de Fourier de tiempo corto (del inglés short-time Fourier transform, STFT) y la transformada Wavelet continua, las cuales se modelaron a través de una CNN, obteniendo resultados de hasta el 89 % para la clasificación de pacientes con EP. Es de aclarar que los autores utilizaron un algoritmo de optimización Bayesiana para encontrar los híper-parámetros de la CNN, lo que puede ocasionar que los resultados sean un poco optimistas o al menos, difíciles de reproducir. En [12] los autores propusieron una estrategia basada en aprendizaje multitarea y el modelo presentado en [8], con el objetivo de mejorar la generalización de las características aprendidas por la CNN y al mismo

tiempo evaluar diferentes déficits del habla en pacientes con EP. En total se tomaron 11 aspectos del habla y concluyeron que el enfoque propuesto mejora la generalización de la red convolucional, teniendo resultados de hasta un 4 % mejor en relación con las redes entrenadas individualmente.

Por otro lado, en [13], los autores propusieron una clasificación de pacientes con EP vs. Controles sanos a través de CNNs usando aprendizaje por transferencia y técnicas de aumento de datos en escritura manuscrita. Para esto utilizaron una base de datos de 576 muestras de 72 personas (36 pacientes con EP y 36 controles sanos), donde cada uno realizó 8 tareas capturadas en una tableta digitalizadora. Los autores tomaron como modelos base CNNs entrenadas con el conjunto de datos de ImageNet (1.2 millones de imágenes y 1000 clases), y el conjunto de datos del MNIST (0.6 millones de imágenes y 10 clases), luego realizaron transferencia de aprendizaje a partir del congelamiento total de las capas y un ajuste fino en las capas de salida. El mejor desempeño se logro con un 98.28 % de precisión usando el método de ajuste fino y tomando de modelo base el conjunto de datos de ImageNet. En [14] presentan un breve resumen de los métodos de aprendizaje por transferencia más usados, particularmente dentro del paradigma moderno del aprendizaje profundo, se realizó un enfoque principal a aplicaciones en el procesamiento del habla y el lenguaje. Los autores concluyeron que el aprendizaje por transferencia se vuelve mucho más fácil y más efectivo con los mapas de características aprendidos por los modelos profundos, y la transferencia puede llevarse a cabo no solo entre distribuciones de datos y tipos de datos, sino también entre estructuras de modelos (modelos monolingües). En [15] se propuso transferir las representaciones de imágenes aprendidas en una CNN con un conjunto de datos grandes (ImageNet), a otras tareas de reconocimiento visual con datos de entrenamiento limitados (Pascal VOC). Los autores analizaron el rendimiento del aprendizaje por transferencia y mostraron mejoras significativas en las tareas de clasificación de objetos, superando en un 8 % el rendimiento de la CNN entrenada solo con Pascal VOC.

1.3. Hipótesis

Es posible mejorar la clasificación de pacientes con enfermedad de Parkinson y controles sanos a partir de aprendizaje por transferencia en datos monolingües.

1.4. Objetivos

1.4.1. Objetivo general

Implementar y evaluar el método de aprendizaje por transferencia en CNNs para tres diferentes idiomas, con el fin de apoyar el diagnóstico y monitorear pacientes con EP.

1.4.2. Objetivos específicos

1. Implementar algoritmos de preprocesamiento y segmentación de señales de voz, para la extracción de transiciones onset y offset.

2. Diseñar y entrenar CNNs de topología ResNet para diferentes idiomas a partir de representaciones tiempo-frecuencia de las transiciones.

3. Implementar la técnica de aprendizaje por transferencia en los modelos entrenados de CNNs para la evaluación y el monitoreo de pacientes con EP.

4. Evaluar y comparar el desempeño de las CNNs entrenadas en diferentes idiomas y las CNNs implementadas con la técnica de aprendizaje por transferencia.

1.5. Contribución de este trabajo

En este trabajo se propone la clasificación de pacientes con EP en tres idiomas diferentes (Español, Alemán, Checo) utilizando aprendizaje por transferencia en CNNs. El método propuesto está basado en el análisis articulatorio presentado en [9] que hace referencia a las transiciones entre segmentos sonoros y no sonoros. Luego estas tramas serán representadas en un espacio tiempo-frecuencia a partir de la STFT. Con esta representación se realizará el entrenamiento de CNNs con datos monolingües, posteriormente se realizara la transferencia de aprendizaje entre idiomas y se evaluará el desempeño de la red a partir de diferentes medidas de desempeño como eficiencia, sensibilidad y especificidad.

2. Marco teórico

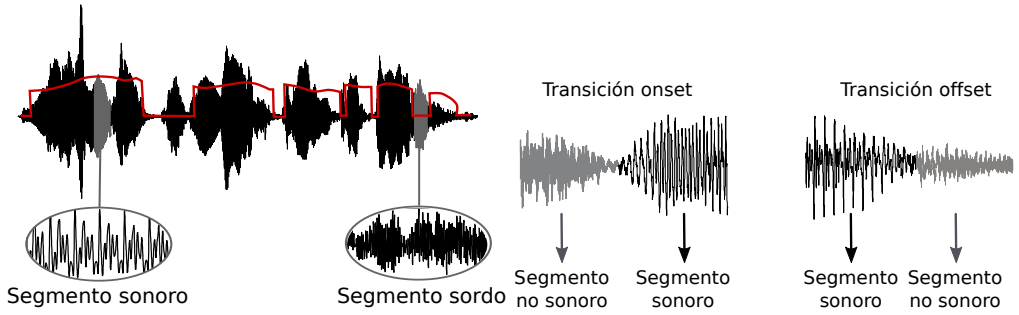
2.1. Análisis articulatorio en señales de voz

El proceso de articulación está relacionado con la modificación de la posición, fuerza y forma de varias extremidades y músculos involucrados en el proceso de la producción del habla, en otras palabras, es el acto de posicionar correctamente los órganos articulatorios para producir un sonido específico durante un tiempo determinado [2]. El análisis articulatorio se puede realizar a partir de vocales sostenidas o habla continua. Cuando se realiza un análisis de vocales sostenidas comúnmente se usan medidas que permiten evaluar la posición de la lengua, mientras que un análisis de habla continua se enfoca principalmente en las transiciones que se presentan en las cuerdas vocales, es decir, la capacidad articulatoria para iniciar y detener la vibración de estas [16]. En este trabajo el análisis se centra en las transiciones de señales de habla continua mediante una representación tiempo-frecuencia, debido a que los pacientes con EP comúnmente producen sonidos anormales en segmentos no sonoros y tienen dificultad para comenzar y/o detener la vibración de los pliegues vocales.

2.2. Detección de transiciones

En señales de voz se puede definir una transición como el cambio entre un segmento sonoro a un segmento no sonoro y viceversa. La característica principal de un segmento sonoro es que son señales quasi-periódicas debido a vibraciones cíclicas de las cuerdas vocales. Por otro lado, en la producción de sonidos sordos no hay vibración de las cuerdas vocales y son señales muy similares al ruido. En este trabajo la identificación de los segmentos sonoros y no sonoros se realiza teniendo en cuenta la presencia de la frecuencia fundamental de la voz en pequeñas tramas de voz [9]. La Figura 1(a) muestra el contorno de la frecuencia fundamental en color rojo de la señal de voz, el cual determina que tipo de segmento es la trama analizada, también es posible observar la quasi-periodicidad del segmento sonoro y la similitud al ruido del segmento sordo.

Para la construcción de las transiciones inicialmente se detecta el cambio entre segmentos, luego se toman 80 ms de la señal hacia la izquierda y 80 ms de la señal a la derecha, formando segmentos de señal con una longitud de 160 ms. A partir de estos segmentos pueden existir 2 tipos de transiciones: **(i)** Transición onset que hace referencia al comienzo de un segmento sonoro a partir un segmento no sonoro. **(ii)** Transición offset que es el cambio de un segmento sonoro a un segmento no sonoro (ver Figura 1(b)).

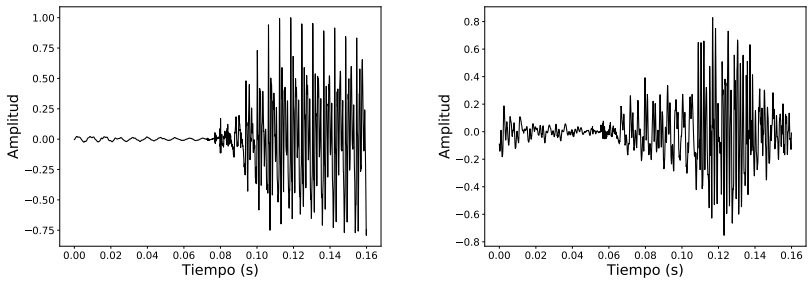


(a) Segmento sonoro y no sonoro. (b) Transición onset y una transición offset.

Figura 1. Segmentación de una señal de audio.[a]

[a]Tomada de: Arias-Vergara, T., Vásquez-Correa, J. C., Orozco-Arroyave, y col.," Speaker models for monitoring Parkinson's disease progression considering different communication channels and acoustic conditions". Speech Communication, 2018, vol. 101, p. 11-25.

En la Figura 2 es posible observar una transición onset de un control sano masculino de 54 años de edad (Figura 2(a)), y una transición onset de un paciente con EP masculino, con 48 años de edad y un nivel UPDRS-III de 9 (Figura 2(b)). Es posible observar que para el control, la transición presenta una mínima oscilación antes de comenzar el segmento sonoro, caso contrario para el paciente con EP, ya que no es tan evidente el comienzo del sonido sonoro debido a grandes oscilaciones en la señal de voz.



(a) Control sano. (b) Paciente con Parkinson.

Figura 2. Transición onset para un control sano y un paciente con EP.

2.3. Características de articulación

Comúnmente la caracterización del sistema articulatorio se realiza a partir de 2 métodos, los MFCCs y la energía de la señal distribuida en la banda

de Bark, los cuales se realizan en las transiciones de la señal (onset-offset), luego a estos métodos se les extrae medidas estadísticas como la media, la desviación estándar, la asimetría y la kurtosis definiendo un vector característico para cada paciente y control sano que finalmente es interpretado por algoritmos de aprendizaje.

Coeficientes cepstrales en las frecuencias de Mel

Los MFCCs son coeficientes para la representación del habla basados en la percepción auditiva humana. Es un método que extrae las componentes de la señal de audio que son buenas para identificar el contenido lingüístico y descartar todo lo demás que transporta información como ruido de fondo, emociones, etc [17]. Los MFCCs se calculan a partir de 5 pasos:

1. Separar la señal en tramas de corta duración.
2. Aplicarle la Transformada de Fourier discreta a cada trama, obteniendo la potencia espectral de la señal.
3. Aplicar un banco de filtros correspondientes a la Escala Mel (ver Figura 3) al espectro obtenido en el paso anterior y sumar las energías en cada uno de ellos.
4. Calcular el logaritmo de todas las energías en cada frecuencia Mel.
5. Aplicarle la transformada de coseno discreta a los logaritmos.

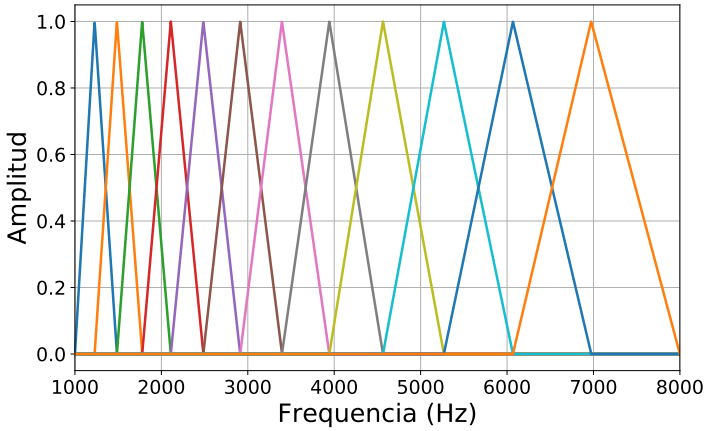


Figura 3. Bandas críticas en frecuencia de Mel.

Energía en la banda de Bark

La escala de Bark es una escala psicoacustica y consiste en un mecanismo de filtrado y traducción de las señales oscilatorias y vibraciones a señales eléctricas interpretables por el sistema nervioso central, en otras palabras, es una escala que representa el rango de frecuencias audibles. En la extracción de características se toma el espectro de las transiciones y se distribuye en 22 bandas criticas siguiendo la escala de Bark. Para frecuencias inferiores a 500 Hz, los anchos de banda de las bandas críticas son constantes a 100 Hz, mientras que para frecuencias medias y altas el incremento es proporcional al logaritmo de frecuencia (ver Ecuación 1) [17].

$$\text{Bark}(f) = 13 \arctan\left(\frac{0{,}76f}{1000}\right) + 3{,}5 \arctan\left(\frac{f}{7500}\right)^2 \tag{1}$$

2.4. Representación tiempo-frecuencia

Para realizar una representación tiempo-frecuencia se utiliza la Transformada de Fourier de Tiempo Corto, la cual es usada para determinar el contenido de frecuencias armónicas y de fase en secciones locales de una señal. El cálculo de la STFT consiste en tomar un determinado número de muestras por medio de una ventana temporal, luego se halla el contenido de frecuencia (Ω) de las muestras puestas en la ventana, y se representan en una gráfica de dos dimensiones (tiempo-frecuencia).

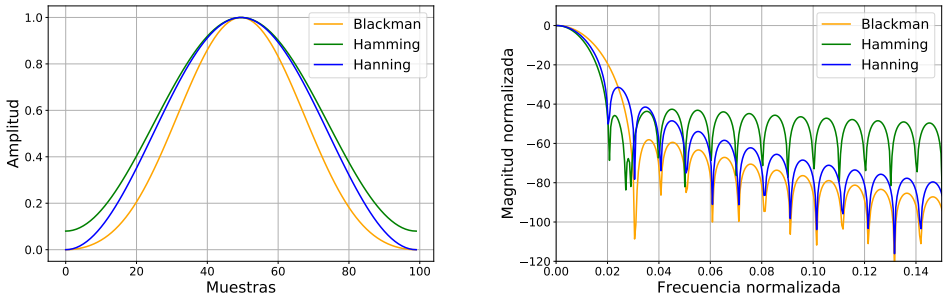
En el caso de señales de audio la información a transformar es dividida en tramas (que usualmente se solapan unas con otras, para reducir irregularidades en la frontera) y a cada una de estas se le realiza una transformación de Fourier a partir de la Ecuación 2.

$$X_m(\Omega) = \sum_{n=-\infty}^{\infty} x(n) f(n-m) e^{-j\Omega n} \tag{2}$$

En la Ecuación 2, $x(n)$ representa la señal de audio y $f(n)$ la ventana utilizada. El índice de tiempo discreto m es normalmente considerado como un tiempo *lento* y usualmente no se expresa con tan alta resolución como con el tiempo n.

En la STFT la ventana constituye un parámetro de gran importancia ya que a través de ésta se puede establecer el grado de resolución tanto de tiempo como de frecuencia que se desee. Si la ventana es muy angosta analizaremos una porción muy pequeña de la señal lo que nos permite tener una

buena resolución en tiempo pero una mala resolución en frecuencia ya que conoceremos sólo una mínima fracción del espectro total de la señal. Por otro lado, si la ventana en muy ancha tendremos una buena resolución en frecuencia pero una mala resolución en tiempo. Entre las ventanas más usadas para realizar este tipo de análisis se encuentran las ventanas Blackman, Hamming, Hanning. (ver Figura 4) [18], [19].



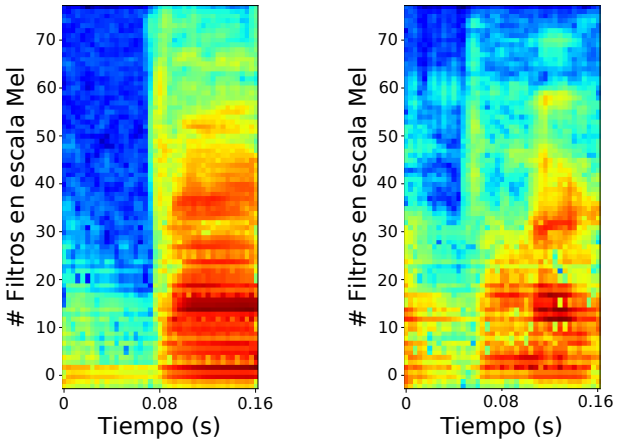
(a) Dominio del tiempo (b) Dominio de la frecuencia

Figura 4. Ventanas Blackman, Hamming, Hanning.

Finalmente la respuesta en magnitud de la STFT es conocida como un espectrograma, el cual es una matriz que muestra la variación del espectro y la energía de la señal para cada una de las tramas a lo largo del tiempo, es común representar la energía de la señal en decibelios (dB) a partir de la Ecuación 3 para una mejor visualización en el espacio de color.

$$\text{Espectrograma}\{x(t)\} = 10 \log_{10} |X_m(\Omega)| \tag{3}$$

En este trabajo se calculan los espectrogramas usando frecuencias en la escala Mel, que tiene como propósito describir el sistema auditivo humano de forma lineal. El banco de filtros usados tiene mayor resolución en baja frecuencia y con mayor concentración alrededor del área percibida por el oído humando, mientras que a altas frecuencias no es tan relevante la información [20] (ver Figura 3). En la Figura 5 es posible observar los espectrogramas correspondientes a las transiciones mostradas en la Figura 2, representados en 80 coeficientes de Mel, una ventana Hanning con un tamaño de 256 muestras que corresponden a 16ms y un solape por ventana de 4ms. Formando finalmente un espectrograma de 80x41, que serán las entradas de la CNN. En esta figura se puede identificar claramente la transición onset del control en el tiempo 0.08, mientras que para el paciente con EP a pesar de que la transición se presenta en el mismo tiempo, ésta no es tan evidente, debido a la presencia de frecuencias durante toda la trama.



(a) Control sano (b) Paciente con EP

Figura 5. STFT de una transición onset para un control sano y un paciente con EP.

2.5. Redes neuronales profundas

En general una red neuronal es un modelo matemático creado para simular la actividad del cerebro humano. Consiste en una gran cantidad de unidades llamadas neuronas artificiales, las cuales se encuentran conectadas entre si para transmitir señales. La información comienza en una capa de entrada, posteriormente atraviesa completamente la red neuronal sometiendo la información a diferentes operaciones y obteniendo valores de salida.

Una Red Neuronal Profunda (del inglés Deep Neural Network, DNN), es una red con un nivel de complejidad mayor. Las DNN utilizan modelos matemáticos sofisticados para procesar datos de manera compleja que nos permiten agrupar y clasificar conjuntos de interés a partir de similitudes o patrones en la información de entrada. Las DNN se distinguen de las redes neuronales convencionales por su profundidad, es decir, por la cantidad de capas a través de las cuales los datos deben atravesar para extraer sus patrones. En las DNN, cada capa oculta se entrena a partir de un conjunto distinto de características basadas en la salida de la capa anterior, cuanto más avance la información en la red neuronal, más complejas serán las características que sus neuronas pueden extraer, ya que agregan y combinan características de capas anteriores. Al final de la red comúnmente se tiene una capa de clasificación, donde son interpretadas las características y se realiza una toma de decisión dependiendo del conjunto de interés [21].

2.5.1. Redes neuronales prealimentadas

La red neuronal prealimentada (del inglés Deep Feedforward Networks) es la base de la mayoría de los modelos de aprendizaje profundo. En esta red, la información fluye hacia adelante, desde la capa de entrada, a través de las capas ocultas y finalmente hasta la capa de salida. No existen conexiones de retroalimentación en las cuales la salida del modelo retorne a la red.

La base de las redes neuronales prealimentadas es el perceptrón, el cual es un modelo similar una neurona humana, que contiene diferentes canales de entrada llamados dendritas y un canal de salida llamado axón. La información es capturada a partir de las dendritas que luego es transmitida al cuerpo de la neurona para obtener una respuesta que es transmitida por el axón. El modelo del perceptrón es mostrado en el recuadro punteado de la Figura 6, este tiene a su entrada un vector $x \in \mathbb{R}^n$, al cual le corresponde un vector de pesos $\omega \in \mathbb{R}^n$. Adicionalmente, se considera un sesgo u que garantiza un nivel mínimo de actividad de la neurona para considerarse como activa. En esta función se interpreta como un regularizador de las señales que se emiten entre neuronas al ponderar las salidas que entran a la neurona. Por último se tiene la función de activación, encargada de activar o apagar la neurona, es decir, si la suma ponderada es superior a cierto umbral, la neurona es disparada y toma el valor de activación (típicamente 1); de forma contraria, toma el valor de desactivación. Este proceso es observado en la Ecuación 4.

$$y = \begin{cases} 1 & \text{si} \quad \sum_{i=0}^{n} \omega_i x_i - u > 0 \\ 0 & \text{si} \quad \sum_{i=0}^{n} \omega_i x_i - u \leq 0 \end{cases} \tag{4}$$

Las redes neuronales típicamente están formadas por una serie de capas de neuronas que están unidas entre si, estas redes comúnmente se les denomina "perceptrón multicapa". Como se observa en la Figura 6, las conexiones del perceptrón multicapa siempre están dirigidas hacia adelante, es decir, las neuronas de una capa se conectan con las neuronas de la siguiente capa, de ahí que reciban el nombre de redes alimentadas hacia adelante, redes prealimentadas o redes "Feedforward" [22], [23].

2.5.2. Función de costo

La función de costo en las redes neuronales es la ecuación que nos determina el error entre el valor estimado y el valor real o etiqueta, con el fin de optimizar los parámetros de la red neuronal. Comúnmente en muchas aplicaciones de las DNN se utiliza una función de costo uniforme como el error cuadrático medio (del inglés Mean Square Error, MSE) mostrada en la Ecuación 5 donde C es la función de costo, N es el número de datos de entre-

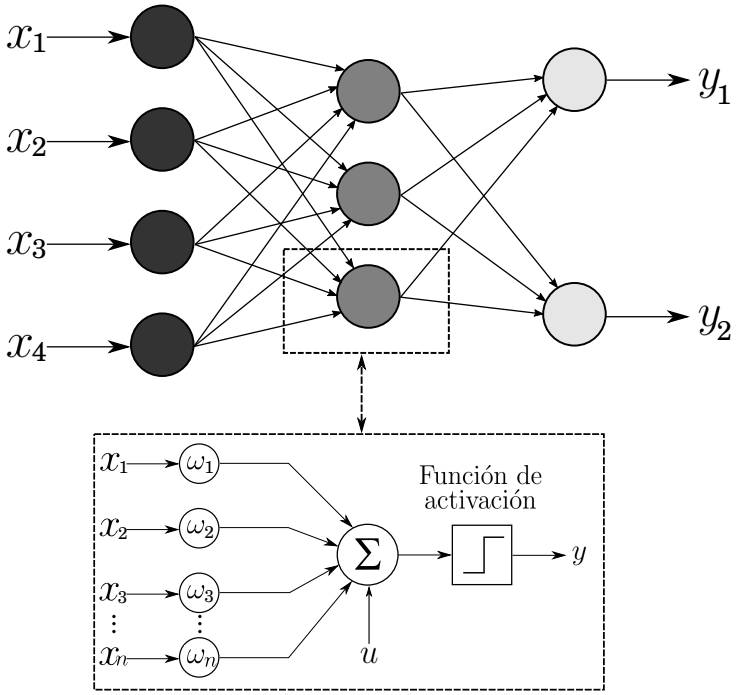


Figura 6. Perceptrón multicapa.

namiento, y_i es el vector de etiquetas reales y y_i^* es el vector de predicciones de la red. MSE como función de costo es usual en problemas de regresión, ya que es sensible a pequeños cambios en los parámetros de la red permitiendo realizar una actualización de los pesos proporcional al error cometido por las neuronas, pero se debe tener en cuenta que este es bastante sensible al ruido de los datos de entrenamiento [21].

$$C(\omega) = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_i^*)^2 \tag{5}$$

Otra función de costo común en el aprendizaje de redes neuronales es la entropía cruzada, donde cada probabilidad estimada por la red neuronal es comparada con el vector de etiquetas reales, luego se calcula una puntuación que penaliza la probabilidad en función a la distancia del valor esperado, en la Ecuación 6 se puede observar esta función de costo donde N corresponde al número de datos de entrenamiento, y_i es el vector de etiquetas reales y p_i son las predicciones de la red neuronal, generalmente interpretadas como probabilidades. Como se puede ver, esta penalización es logarítmica, por lo tanto se tiene una pequeña puntuación para pequeñas diferencias (0.1 o 0.2) y una gran puntuación para una gran diferencia (0.9 o 1.0). Finalmente un

modelo que predice clases perfectamente, tiene una entropía cruzada de 0.0. Comúnmente la entropía cruzada es usada para problemas de clasificación debido a que la predicción converge rápidamente y de una manera más robusta, a pesar de que se debe tratar con el problema del desvanecimiento del gradiente [24].

$$C(\omega) = \frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{6}$$

2.5.3. Gradiente descendente

El método del gradiente descendente, es uno de los algoritmos de optimización más populares en aprendizaje automático, particularmente en el campo de las redes neuronales. Este método define una función la cual es proporcional al error que comete la red en función del conjunto de parámetros de ésta. A partir de este método es posible encontrar la mejor configuración de los pesos que lleve el algoritmo a encontrar un mínimo global de la función de costo.

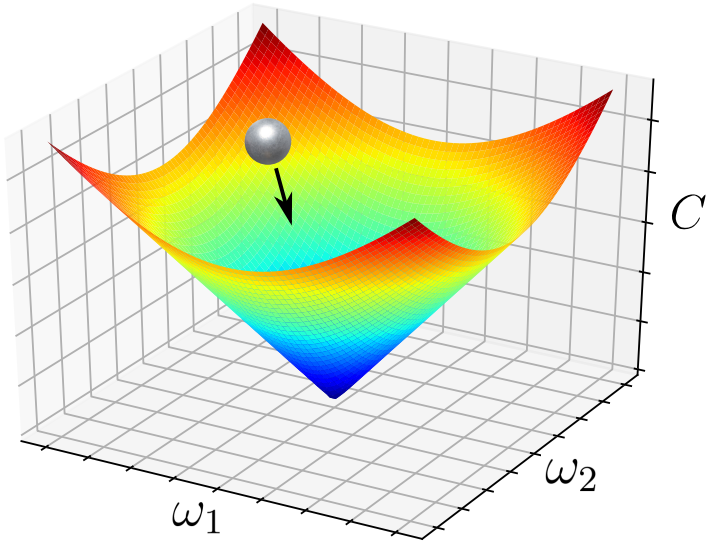


Figura 7. Aprendizaje basado en el método del gradiente descendente.

Como se observa en la Figura 7 lo que se pretende en el método del gradiente descendente es que la función de costo C se minimice a partir de pequeños desplazamientos, específicamente una cantidad $\Delta\omega_1$ en la dirección ω_1, y una cantidad $\Delta\omega_2$ en la dirección ω_2. Matemáticamente C cambia a partir de la Ecuación 7.

$$\Delta C \approx \frac{\partial C}{\partial \omega_1}\Delta\omega_1 + \frac{\partial C}{\partial \omega_2}\Delta\omega_2 \tag{7}$$

De manera general el cambio de la función de costo se puede denotar como $\Delta C \approx \nabla C \cdot \Delta\omega$, donde ∇C está denotado con el Ecuación 8 y $\Delta\omega$ por la Ecuación 9.

$$\nabla C \equiv \left(\frac{\partial C}{\partial \omega_1}, \cdots, \frac{\partial C}{\partial \omega_m}\right) \tag{8}$$

$$\Delta\omega = (\Delta\omega_1, \cdots, \Delta\omega_m)^\intercal \tag{9}$$

Finalmente para encontrar la configuración de pesos óptima mediante el método del gradiente descendente se parte de una configuración de pesos ω_k, donde k es la cantidad de pesos de la red, luego se calcula la dirección de máxima variación de la función de costo $C(\omega)$ dada por los parámetros de la red, el sentido de máxima variación apuntara hacia una colina en la superficie, a continuación se actualizarán los pesos de la red en el sentido opuesto indicado por el gradiente de la función de costo, reduciendo esta y aproximándose en cada iteración al mínimo global de la función. Matemáticamente este proceso es mostrado en la Ecuación 10, donde ω_k representa los pesos iniciales, ω'_k los pesos actualizados, $\frac{\partial C}{\partial \omega_k}$ el sentido de máxima variación y η indica el tamaño del paso en cada iteración, conocida como tasa de aprendizaje [21], [25].

$$\omega_k \to \omega'_k = \omega_k - \eta\frac{\partial C}{\partial \omega_k} \tag{10}$$

2.5.4. Propagación hacia atrás

Propagación hacia atrás (del inglés Back-Propagation), es el método mediante el cual una red neuronal prealimentada ajusta sus parámetros o pesos para aprender una representación interna de la información que está procesando. En otras palabras, es el algoritmo que calcula el sentido de máxima variación en cada capa para posteriormente hacer una actualización de los pesos mediante el gradiente descendente.

Este método se basa en la propagación hacia adelante que realizó la red, es decir, se aplica una entrada x_n a las neuronas iniciales, luego se propaga la información por las distintas capas de la red hasta generar una salida, la cual es comparada con la etiqueta real a partir de la función de costo que determina el error cometido por cada parámetro de la red [21].

Inicialmente partimos de la Ecuación 4, donde se tiene la suma ponderada de los pesos multiplicado por la entrada menos un sesgo, todo esto lo denotamos como z, que luego es pasada por una función de activación denotada como a la cual es propia de la topología de la red, por último este resultado es llevado a la función de costo. Este proceso es resumido en la Ecuación 11, donde L indica el número de capas de la red.

$$C[a(z^L)] = C[a(\omega^L x + u^L)] \tag{11}$$

A partir de esta composición de funciones se realiza la derivada de la Ecuación 11 respecto a ω^L mediante la regla de la cadena, con el fin de hallar el sentido de máxima variación para la capa L.

$$\frac{\partial C}{\partial \omega^L} = \frac{\partial C}{\partial a^L} \cdot \frac{\partial a^L}{\partial z^L} \cdot \frac{\partial z^L}{\partial \omega^L} \tag{12}$$

Los dos primeros términos de la Ecuación 12 hacen referencia a como varia el error de la función de costo cuando hay un cambio en la sumatoria de las neuronas, esta definición es conocida como el error imputado de las neuronas y es denotado como δ^L. Por lo tanto se tiene que:

$$\frac{\partial C}{\partial a^L} \cdot \frac{\partial a^L}{\partial z^L} = \delta^L \tag{13}$$

El último término de la Ecuación 12 es la representación de como varía la suma ponderada z^L con respecto a los pesos de la red, teniendo en cuenta que las entradas de esta capa son las salidas de la capa anterior, denotamos como a_i^{L-1} la salida de la capa $(L-1)$, por lo tanto solucionando la derivada parcial se obtiene que:

$$z^L = \sum_{i=0}^{n} \omega_i^L a_i^{L-1} + u^L \tag{14}$$

$$\frac{\partial z^L}{\partial \omega^L} = a_i^{L-1} \tag{15}$$

Finalmente, tomando las Ecuaciones 13, 15 y reemplazándolas en 12 tenemos que:

$$\frac{\partial C}{\partial \omega^L} = \delta^L \cdot a_i^{L-1} \tag{16}$$

Donde se puede hallar la derivada parcial de la función de costo con respecto a los pesos de esta capa, ya que es posible encontrar el error imputado que depende de la derivada parcial de la función de costo respecto a la función

de activación y la derivada parcial de la función de activación con respecto a la suma ponderada.

Ahora continuando con el algoritmo y realizando el mismo análisis en la capa anterior, es decir, en la capa $(L-1)$, se tiene que:

$$\frac{\partial C}{\partial \omega^{L-1}} = \frac{\partial C}{\partial a^{L}} \cdot \frac{\partial a^{L}}{\partial z^{L}} \cdot \frac{\partial z^{L}}{\partial a^{L-1}} \cdot \frac{\partial a^{L-1}}{\partial z^{L-1}} \cdot \frac{\partial z^{L-1}}{\partial \omega^{L-1}} \tag{17}$$

$$\frac{\partial C}{\partial \omega^{L-1}} = \delta^{L} \cdot \frac{\partial z^{L}}{\partial a^{L-1}} \cdot \frac{\partial a^{L-1}}{\partial z^{L-1}} \cdot a^{L-2} \tag{18}$$

En la Ecuación 18 podemos observar que se tienen 2 términos que son desconocidos, en este caso el término $\frac{\partial z^{L}}{\partial a^{L-1}}$ hace referencia a los pesos que conecta ambas capas, es decir, ω^{L}. Por otro lado $\frac{\partial a^{L-1}}{\partial z^{L-1}}$ es la derivada parcial de la función de activación con respecto a la suma ponderada de las neuronas, la cual es propia de la topología de la red con la que se este trabajando. Por lo tanto es posible conocer el gradiente de la función de costo para la capa $(L-1)$ a partir de la Ecuación 19.

$$\frac{\partial C}{\partial \omega^{L-1}} = \delta^{L} \cdot \omega^{L} \cdot \frac{\partial a^{L-1}}{\partial z^{L-1}} \cdot a^{L-2} \tag{19}$$

En resumen el algoritmo de Back-Propagation funciona a partir de los siguientes 3 pasos deducidos de la explicación anterior, comenzado desde la ultima capa y realizando el mismo proceso secuencialmente hasta terminar en la capa inicial:

1. Cómputo del error de la ultima capa (L):

$$\delta^{L} = \frac{\partial C}{\partial a^{L}} \cdot \frac{\partial a^{L}}{\partial z^{L}} \tag{20}$$

2. Retro-propagamos el error a la capa anterior $(L-1)$:

$$\delta^{L-1} = \omega^{L} \cdot \delta^{L} \cdot \frac{\partial a^{L-1}}{\partial z^{L-1}} \tag{21}$$

3. Calculamos las derivadas de la capa usando el error:

$$\frac{\partial C}{\partial \omega^{L-1}} = \delta^{L-1} \cdot a^{L-2} \tag{22}$$

2.5.5. Regularización

Uno de los grandes problemas en el aprendizaje de las redes neuronales profundas es cómo hacer para que el algoritmo funcione bien tanto para los datos de entrenamiento, como para nuevas entradas. La regularización es el área encargada de reducir el error de generalización del algoritmo de aprendizaje, es decir, evita el sobre-ajuste del modelo.

En el ámbito del aprendizaje profundo el espacio de soluciones de las redes neuronales profundas es bastante amplio, en otras palabras, este método de aprendizaje puede adaptarse con gran facilidad a nuestros datos de entrenamiento, provocando un excelente rendimiento con los datos de entrenamiento pero un mal desempeño con los datos de prueba. Lo que indica que si no se utiliza una contra-medida seria bastante complejo entrenar el algoritmo. En este trabajo nos enfocamos en 3 diferentes métodos de regularización los cuales son explicados a continuación.

Regularización L^2

Este tipo de regularización es uno de los más comunes en el ámbito del aprendizaje profundo, también conocido como regularización de Tikhonov. Este método es una forma específica de regularizar la función de costo, adicionando el término $\frac{1}{2}\|\omega\|_2^2$, que significa la norma Euclidiana al cuadrado de los pesos de la red, además se agrega un parámetro adicional λ que multiplica la norma Euclidiana y nos permite tener un control del nivel de regularización. Por lo tanto la Ecuación 5 que representa la función de costo es modifica de la siguiente manera:

$$C(\omega)' = C(\omega) + \frac{1}{2}\lambda\|\omega\|_2^2 \tag{23}$$

La actualización de los pesos también se ve modificada, es decir, la Ecuación 10 se convierte en:

$$\omega_k \to \omega_k' = \omega_k - \eta\frac{\partial C}{\partial \omega_k} - \lambda\omega_k \tag{24}$$

En la Ecuación 24 se puede observar que cada vez que se realiza la actualización de los pesos, se le resta $\lambda\omega_k$, esto le da a los pesos una tendencia a decaer a cero, de ahí el nombre conocido como caída de peso (del inglés weight decay).

Dropout

Es una técnica de regularización que desactiva un número de neuronas aleatoriamente. En cada iteración de entrenamiento el dropout desactiva diferentes neuronas, las cuales no se tienen en cuenta para el Forward-Propagation ni para el Back-Propagation evitando que las neuronas sean dependientes de otras cercanas. Este método ayuda con el sobre-ajuste de la red ya que comúnmente las neuronas aprenden patrones los cuales se adaptan al conjunto de entrenamiento. Con dropout la dependencia entre neuronas es menor, de tal manera que éstas deban trabajar de manera solitaria y no depender tanto de sus vecinas. El parámetro del dropout nos indica la probabilidad de desactivar las neuronas y toma valores entre 0 y 1, donde un valor cercano a 0 nos indica un menor número de neuronas desactivas y un valor cercano a 1 gran cantidad de neuronas apagadas [26].

Early stopping

Comúnmente en el entrenamiento de los modelos se observa que el error de entrenamiento disminuye constantemente con el tiempo, pero en ciertos casos el error del conjunto de validación comienza a aumentar, esto significa que se esta perdiendo la generalización del modelo y éste se está sobre-ajustando a los datos de entrenamiento. Esto implica que la mejor configuración del modelo se encuentra cuando se tiene el menor error en el conjunto de validación. La idea general del algoritmo de early stopping es que cada vez que se mejore el error en el conjunto de validación, se almacena una copia de los parámetros del modelo, cuando este error comienza a aumentar el algoritmo de entrenamiento se detiene luego de un número de iteraciones específico, retornando los parámetros de la ultima configuración almacenada (ver Figura 8). Este procedimiento es quizás uno de los métodos más antiguos y más utilizados de regularización en redes neuronales. En este método de regularización el parámetro a controlar es el número de iteraciones que el algoritmo debe esperar para terminar el proceso de entrenamiento.

2.6. Redes neuronales convolucionales

Las CNNs son muy similares a las redes neuronales convencionales (como el perceptrón multicapa), las cuales están compuestas por neuronas con sus respectivos pesos que son calculados en una etapa de entrenamiento. Una de las principales características de las CNNs es que permiten reducir la cantidad de parámetros a partir de capas de pooling, lo que conlleva a consumir menos tiempo computacional, sin dejar atrás la robustez del sistema. Una

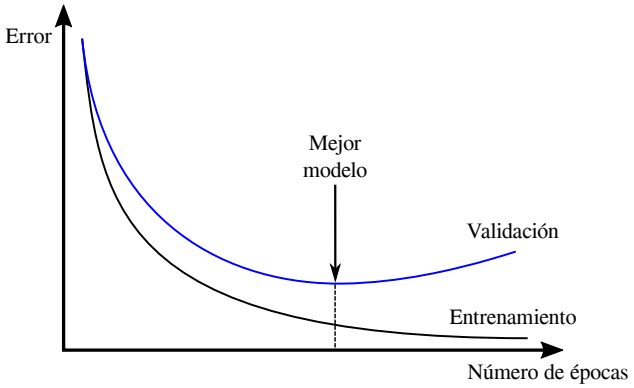


Figura 8. Early stopping.

CNN típicamente consta de 3 etapas (ver Figura 9): en la primera etapa, la capa realiza varias convoluciones en paralelo para producir un conjunto de activaciones lineales. En la segunda etapa, utilizamos una función de reducción o pooling para modificar la salida de la capa, la cual reduce la cantidad de parámetros y obtiene las características más relevantes. Y por último una capa de clasificación totalmente conectada, la cual es la encargada de dar el resultado final de la red [21], [27].

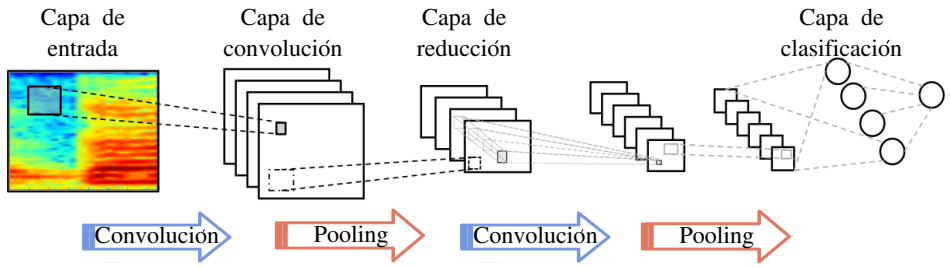


Figura 9. Estructura típica de una CNN.

2.6.1. Etapa de convolución

La convolución es una operación típicamente utilizada para caracterizar sistemas lineales. Las CNN son redes que utilizan dicha operación en lugar de la multiplicación general de matrices en al menos una de sus capas. La convolución es típicamente denotada con un asterisco (Ecuación 25).

$$s(t) = (x * w)(t) = \int x(a)w(t-a)da \tag{25}$$

Para tiempo discreto, la convolución se obtiene de acuerdo con la Ecuación 26.

$$s(t) = (x * w)(t) = \sum_{a=-\infty}^{\infty} x(a)w(t-a) \tag{26}$$

En la terminología de las CNN, el primer argumento (x) de la convolución a menudo se conoce como la entrada, y el segundo argumento (w) como el núcleo o kernel, y la salida es conocida como el mapa de características. Comúnmente en aplicaciones de aprendizaje de máquina la entrada suele ser una matriz de datos multidimensional, y el kernel suele ser una matriz multidimensional de parámetros que son adaptados por el algoritmo de aprendizaje.

A menudo utilizamos convoluciones en más de un eje a la vez. Por ejemplo, si usamos una imagen bidimensional I como nuestra entrada, probablemente también queramos usar un kernel bidimensional W, por lo tanto:

$$S(i,j) = (I * K)(i,j) = \sum_{m} \sum_{n} I(m,n)W(i-m, j-n) \tag{27}$$

En este caso la convolución discreta se puede ver como una multiplicación de matrices. La matriz de entrada tiene varias restricciones debido a que comúnmente el kernel suele ser mucho más pequeño que la imagen de entrada (ver Figura 10). Esto permite modelar de forma consecutiva pequeñas piezas de información y luego combinar información en capas posteriores de la red. Una manera de entender las CNNs es que la primera capa intentará detectar los bordes y establecer patrones de detección de bordes. Luego, las capas posteriores tratarán de combinarlos en formas más simples y, finalmente, en patrones de las diferentes posiciones de los objetos, iluminación, escalas, etc [21].

2.6.2. Etapa de reducción o pooling

La capa de reducción o pooling se ubica generalmente después de la capa convolucional. Su utilidad principal radica en la reducción de las dimensiones espaciales de la capa de entrada a partir de un resumen estadístico de las salidas más cercanas en la capa. La operación realizada por esta capa también se llama submuestreo, ya que la reducción de tamaño conduce también a la pérdida de información. Sin embargo, una pérdida de este tipo puede ser beneficioso para la red por dos razones:

1. La disminución en el tamaño conduce a una menor sobrecarga de cálculo para las próximas capas de la red.

2. Reduce el sobre-ajuste.

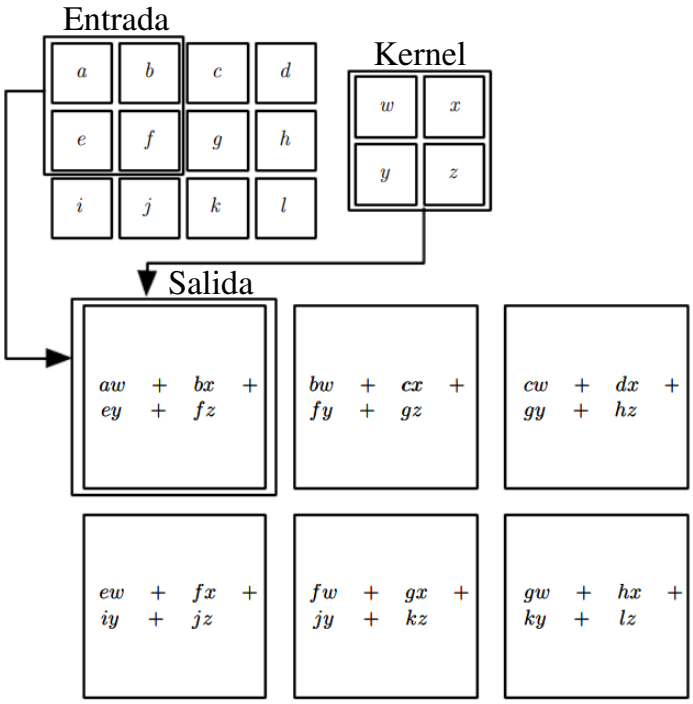


Figura 10. Ejemplo de una convolución bidimensional. [a]

[a]Tomada de: Goodfellow, I. (2016). Deep learning. MIT press.

Una de las operaciones más comunes en esta capa es llamada Max pooling (ver Figura 11), es un método que reporta el máximo valor de salida a partir de sus vecinos más cercanos en un conjunto rectangular. Otra función bastante utilizada en la capa de pooling hace referencia al promedio de un vecindario rectangular, la norma L^2 de un vecindario rectangular o un promedio ponderado basado en la distancia desde el píxel central [21], [27].

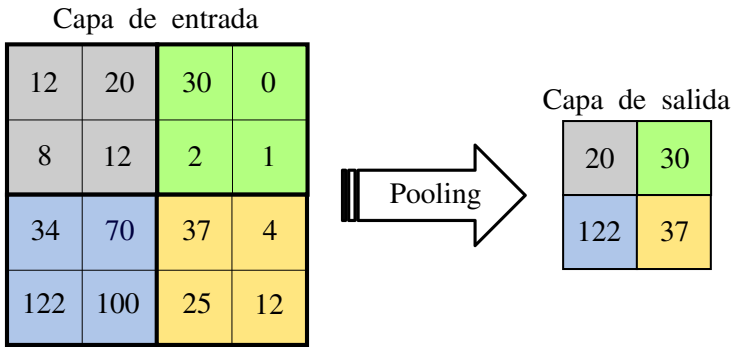


Figura 11. Capa Pooling usando el método max pooling.

2.6.3. Etapa de clasificación

Al final de las capas de convolución y de pooling, las redes utilizan generalmente capas completamente conectados en la que cada píxel se considera como una neurona separada al igual que en una red neuronal regular. Esta última capa clasificadora tendrá tantas neuronas como el número de clases que se debe predecir.

2.6.4. Conexiones residuales (ResNet)

Una de las topologías que ha sobresalido en el aprendizaje profundo es la ResNet para darle solución a la problemática del desvanecimiento del gradiente, cuyo descenso dado por la minimización de la función de error, se reduce exponencialmente a través de la propagación de las capas anteriores, dificultando el aprendizaje de capas profundas [28]. ResNet permite tener un aprendizaje continuo del gradiente, ya que este agrega información adicional a través de la operación adición. Otra gran ventaja de esta topología es que permite construir redes de innumerables capas (potencialmente más de mil). La estructura de adición que conlleva esta metodología no introduce parámetros adicionales ni complejidad de cómputo permitiendo sobresalir ante otras arquitecturas como VGGNet y GoogLeNet, obteniendo un mejor desempeño a igual o menor costo computacional [29].

La topología ResNet está compuesta de una serie de capas y un mapeo de identidad que agrega una entrada de bloque a la salida, es decir, en lugar de tratar de aprender a partir de un mapeo directo de x y con una función $H(x)$, se puede definir una función residual $F(x) = H(x) - x$, que se puede reescribir $H(x) = F(x) + x$; donde $F(x)$ representa las capas apiladas de la red neuronal y x la función de identidad, la cual es llevada hasta la salida del bloque (ver Figura 12).

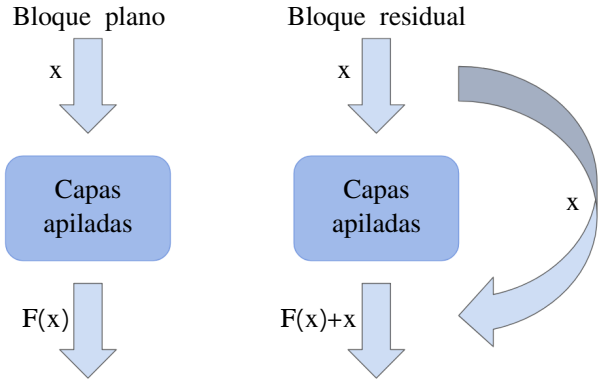


Figura 12. Mapeo de identidad en bloques residuales.

Si el mapeo de identidad es óptimo, podemos fácilmente llevar los residuos a cero ($F(x) = 0$) en vez de ajustar un mapeo de identidad (x, entrada = salida) a través de la pila de capas de la red. En otras palabras, es más fácil encontrar una solución como $F(x) = 0$ en lugar de $F(x) = x$ usando la pila de capas no lineales de la CNN como función [29].

2.7. Aprendizaje por transferencia

La idea inicial del aprendizaje por transferencia es reutilizar la experiencia obtenida para mejorar el aprendizaje de nuevos modelos. En el aprendizaje por transferencia se puede aprovechar el conocimiento (características, pesos, etc.) de modelos previamente creados para entrenar modelos nuevos e incluso abordar problemas de modelos con pequeñas cantidades de datos, a diferencia del aprendizaje tradicional que está aislado y se basa exclusivamente en tareas específicas, conjuntos de datos y entrenamiento sobre modelos separados, donde no se retiene ningún tipo de conocimiento que pueda ser transferido de un modelo a otro (ver Figura 13) [14], [30], [31].

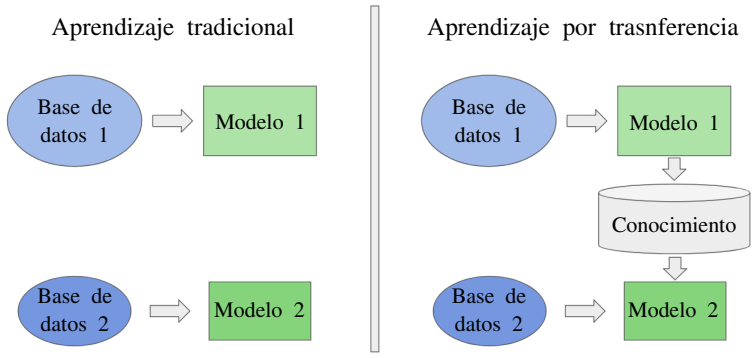


Figura 13. Comparación entre métodos de aprendizaje tradicional y métodos de aprendizaje por transferencia.

El aprendizaje profundo ha progresado considerablemente en los últimos años. Esto ha permitido abordar problemas complejos y producir buenos resultados. Sin embargo, el tiempo de entrenamiento y la cantidad de datos requeridos para tales sistemas de aprendizaje son bastante grandes. Existen varias redes de aprendizaje profundo con un rendimiento bastante bueno (incluso mejor que el desempeño humano) que se han desarrollado y probado en diferentes tópicos obteniendo un gran desarrollo en aplicaciones de visión por computadora y el procesamiento de lenguaje natural. Estos modelos formados previamente son la base del aprendizaje por transferencia en el contexto del aprendizaje profundo [30], [31].

Las señales de voz son señales no estacionarias y cambian enormemente de acuerdo con un gran número de factores (idioma, género, hablante, canal, ambiente, emoción, etc.). Tratar con estas variaciones es el desafío principal de la investigación del procesamiento del habla, y el aprendizaje por transferencia es una herramienta importante para dar solución a esta problemática. A continuación se da una breve explicación de los campos más destacados a partir del aprendizaje por transferencia en el habla. [14].

Transferencia entre idiomas

Es natural creer que algunos patrones del habla y la acústica son compartidos en todos los idiomas. Por ejemplo, muchas consonantes y vocales son similares, este intercambio entre idiomas se ha utilizado explícita o implícitamente para mejorar la solidez estadística en condiciones multilingües, y ha brindado mejores modelos que el entrenamiento con datos monolingües. Su idea básica se centra en que las características aprendidas por los modelos tienden a ser independientes del lenguaje en las capas bajas y más dependientes del lenguaje en las capas superiores [14].

Transferencia entre hablantes

La adaptación del hablante es otro dominio en el que el aprendizaje por transferencia ha sobresalido. En el paradigma de modelos estadísticos paramétricos, por ejemplo, modelos Gaussianos o modelos de mezclas Gaussianas, estimación de maximo a posteriori y la regresión lineal de máxima verosimilitud (MLLR) son los métodos más exitosos para adaptar un modelo a un hablante específico. Su idea principal se basa en tener un modelo ya entrenado que permita adaptar fácilmente el hablante de interés con ciertas restricciones ya establecidas [14].

Transferencia de modelos

Un progreso reciente en el aprendizaje por transferencia es aprender un nuevo modelo (denotado modelo infantil) a partir de un modelo existente (modelo maestro), que se conoce como transferencia del modelo. La idea principal es que el modelo maestro aprenda un rico conocimiento de los datos de entrenamiento y este conocimiento puede usarse para guiar el entrenamiento de modelos infantiles que son simples y por lo tanto no pueden aprender muchos detalles sin la guía del maestro [14].

2.8. Medidas de desempeño

Una medida de desempeño o medida de rendimiento es una medición para conocer la efectividad y la eficiencia de un modelo. Para aplicaciones relacionadas con el aprendizaje de máquinas es común tener medidas como la eficiencia, sensibilidad y especificidad [32].

Para una mejor interpretación de estos términos, inicialmente debemos definir la matriz de confusión, la cuál evalúa el desempeño del sistema dependiendo del número de aciertos y fallos en la etapa de clasificación de los datos de prueba. La Tabla 1 muestra una matriz de confusión para un sistema de clasificación bi-clase, de acuerdo a esta tabla se definen los siguientes términos:

- ✓ **Verdaderos positivos** (del inglés True positive, TP): Número de personas correctamente identificadas como pacientes con EP.
- ✓ **Verdaderos negativos** (del inglés True negative, TN): Número de personas correctamente identificadas como controles sanos.
- ✓ **Falsos positivos** (del inglés False positive, FP): Número de personas incorrectamente identificadas como pacientes con EP.
- ✓ **Falsos negativos** (del inglés False negative, TP): Número de personas incorrectamente identificadas como controles sanos.

Tabla 1. Matriz de confusión. EP: Enfermedad de Parkinson, CS: Control Sano

		Clase estimada	
		EP	**CS**
Clase Verdadera	**EP**	TP	FN
	CS	FP	TN

Eficiencia

La eficiencia (Ecuación 28), es la proporción del sistema de clasificar correctamente pacientes con EP y controles sanos.

$$\text{Eficiencia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{28}$$

Sensibilidad

La sensibilidad (Ecuación 29), es la la capacidad del sistema para detectar pacientes con EP.

$$\text{Sensibilidad} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{29}$$

Especificidad

La especificidad (Ecuación 30), es la la capacidad del sistema para detectar personas sanas.

$$\text{Especificidad} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{30}$$

Curva ROC

La curva ROC (del inglés Receiver Operating Characteristic o Característica Operativa del Receptor) es una representación gráfica del rendimiento del sistema a partir de la tasa de verdaderos positivos (del inglés True Positive Rate, TPR) o sensibilidad y la tasa de falsos positivos (del inglés False Positive Rate, FPR) o (1 – especificidad). Para la construcción de la curva ROC podemos observar la Figura 14, donde el TPR define cuántos resultados positivos correctos ocurren entre todas las muestras positivas a medida que se desplaza un umbral de decisión, denotado por la linea vertical en la gráfica de la izquierda. Por otro lado, el FPR define cuántos resultados positivos incorrectos ocurren entre todas las muestras negativas durante la prueba durante el desplazamiento del umbral. La línea punteada en la gráfica de la derecha hace referencia a los aciertos al azar, es decir, una decisión aleatoria del clasificador. Finalmente, la curva ROC nos permite comparar diferentes modelos y seleccionar los sistemas mas eficientes para la clasificación de la clase de interés.

Coeficiente kappa de Cohen

El coeficiente kappa de Cohen es una medida estadística que refleja la concordancia entre dos evaluadores, pero a su vez mide el grado de acuerdo que puede ser atribuido al azar. El coeficiente kappa puede tomar valores entre -1 y 1, mientras más cercano a 1, el grado de concordancia de los evaluadores es mayor, por el contrario, mientras más cercano a -1, mayor es el

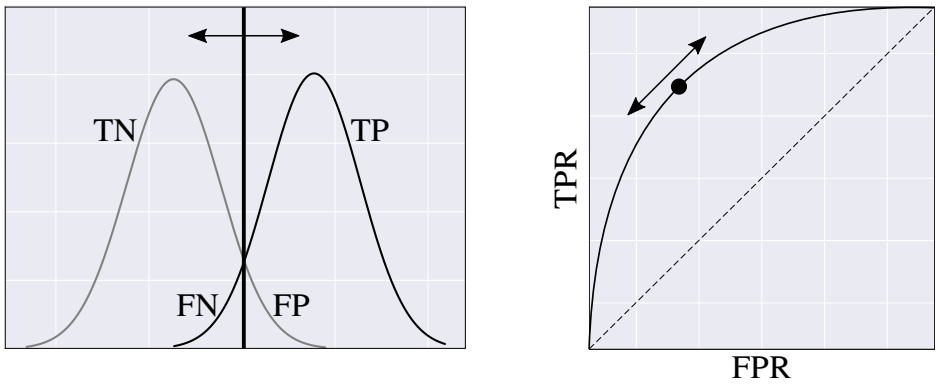


Figura 14. Construcción de la Curva ROC.

grado de discordancia entre los observadores. Un valor de $\kappa = 0$ refleja que la concordancia observada es precisamente la que se espera a causa exclusiva del azar. En la Ecuación 31 es posible observar este coeficiente donde p_o representa la proporción de acuerdos observados y p_ϵ la proporción de acuerdos por azar [33].

$$\kappa = \frac{p_o - p_\epsilon}{1 - p_\epsilon} \tag{31}$$

3. Metodología

3.1. Bases de datos

Tres conjuntos de bases de datos en diferentes idiomas fueron consideradas en este estudio, una base de datos en Español, una en Alemán y una en Checo; cada una formada por pacientes con EP y controles sanos, los cuales fueron grabados en condiciones acústicas controladas. Finalmente todas las grabaciones fueron re-muestreadas a la mínima frecuencia de muestro, la cual corresponde a 16 kHz.

Español

PC-GITA es una base de datos que contiene grabaciones de voz de 50 pacientes con EP y 50 controles sanos muestreados a 44.1 kHz [34]. Todos los participantes son colombianos nativos, balanceados en edad y género, la Tabla 2 muestra información adicional de los hablantes de esta base de datos. Los participantes realizaron diferentes tareas de habla, incluyendo la repetición rápida de silabas como /pa-ta-ka/, /pa-ka-ta/, /pe-ta-ka/, /pa/, /ta/, /ka/, frases aisladas, un texto leído y un monólogo. Todos los pacientes fueron grabados en estado ON, es decir, no mas de tres horas después de tomar su medicación.

Tabla 2. Información general de los hablantes de PC-GITA. μ: media, σ: desviación estándar.

	Pacientes con Parkinson		Controles Sanos	
	Hombres	Mujeres	Hombres	Mujeres
Número de personas	25	25	25	25
Edad ($\mu \pm \sigma$)	61.3 ± 11.4	60.7 ± 7.3	60.5 ± 11.6	61.4 ± 7.0
Rango de edad	33 – 81	49 – 75	31 – 89	49 – 76
Tiempo de diagnóstico ($\mu \pm \sigma$)	8.7 ± 5.8	12.6 ± 11.6		
MDS-UPDRS-III ($\mu \pm \sigma$)	37.8 ± 22.1	37.6 ± 14.1		
Rango de MDS-UPDRS-III	6 – 93	19 – 71		

Alemán

Este corpus está formado por 176 hablantes alemanes nativos, divididos en 88 pacientes con EP, y 88 personas sanas [35]. Los detalles de la base de datos puede ser observada en la Tabla 3. Las grabaciones de voz alemanas fueron capturadas a una frecuencia de muestro de 16 kHz y con los pacientes en estado ON. Todos las personas realizaron tareas de habla como la repetición

rápida de las silabas /pa-ta-ka/, oraciones aisladas, un texto leído y por último un monólogo.

Tabla 3. Información general de los hablantes de la base de datos alemana. μ: media, σ: desviación estándar.

	Pacientes con Parkinson		Controles Sanos	
	Hombres	Mujeres	Hombres	Mujeres
Número de personas	47	41	44	44
Edad ($\mu \pm \sigma$)	66.7 ± 8.4	66.2 ± 9.7	63.8 ± 12.7	62.6 ± 15.2
Rango de edad	44 – 82	42 – 84	26 – 83	28 – 85
Tiempo de diagnóstico ($\mu \pm \sigma$)	7.0 ± 5.5	7.1 ± 6.2		
MDS-UPDRS-III ($\mu \pm \sigma$)	22.1 ± 9.9	23.3 ± 12.0		
Rango de MDS-UPDRS-III	5 – 43	6 – 5		

Checo

Por último la base de datos checa está formada por 50 pacientes con EP y 49 controles sanos capturados a una frecuencia de muestro de 48 kHz [36]. Todos los participantes son checos nativos y realizaron tareas como la repetición rápida de las silabas /pa-ta-ka/, la lectura de un texto y un monólogo. La Tabla 4 muestra mayor información de los participantes en esta base de datos.

Tabla 4. Información general de los hablantes de la base de datos Checa. μ: media, σ: desviación estándar.

	Pacientes con Parkinson		Controles Sanos	
	Hombres	Mujeres	Hombres	Mujeres
Número de personas	30	20	30	19
Edad ($\mu \pm \sigma$)	65.3 ± 9.6	60.1 ± 8.7	60.3 ± 11.5	63.5 ± 11.1
Rango de edad	43 – 82	41 – 72	41 – 77	40 – 79
Tiempo de diagnóstico ($\mu \pm \sigma$)	6.7 ± 4.5	6.8 ± 5.2		
MDS-UPDRS-III ($\mu \pm \sigma$)	21.4 ± 11.5	18.1 ± 9.7		
Rango de MDS-UPDRS-III	4 – 54	6 – 38		

3.2. Validación cruzada

La validación cruzada es una técnica usada comúnmente en el aprendizaje de máquina, este método garantiza la independencia entre los datos de entrenamiento y los datos de prueba, permitiendo reportar resultados menos optimistas y mas cercanos a la realidad.

En este trabajo se utilizó una validación cruzada de K particiones (del inglés K-fold cross-validation) donde los datos son divididos en K subconjuntos, uno es utilizado como dato de prueba y los otros (K-1) como datos de entrenamiento. Este proceso es repetido durante K veces, garantizando que todos los datos hayan sido utilizados como datos de prueba sin repetir ninguno. Finalmente la media aritmética y la desviación estándar son calculadas a las medidas de desempeño con el fin de obtener un único resultado. Cabe aclarar que la validación implementada en este trabajo es independiente de hablante, es decir, garantiza que las grabaciones de una misma persona no estarán en los datos de entrenamiento y en los datos de prueba en la misma partición. Debido a que cada participante tiene varias grabaciones por las diferentes tareas analizadas, y además para cada tarea se tienen diferentes transiciones. La clasificación de cada persona se realiza de acuerdo con una regla de decisión por mayoría, usando la moda de la clase predicha para todos los espectrogramas de cada hablante.

3.3. Experimentos

Para realizar la comparativa y observar si el aprendizaje por transferencia entre idiomas puede mejorar el apoyo del diagnóstico de la EP se plantearon 3 escenarios diferentes: **(i)** Clasificación con máquinas de soporte vectorial a partir de características articulatorias clásicas, para ser usado como referencia. **(ii)** Entrenamiento y clasificación de CNNs con datos monolingües. **(iii)** Clasificación a partir de aprendizaje por transferencia entre idiomas. Por último se realizó un experimento para evaluar el estado de severidad de los pacientes a partir de un clasificador multiclase. En cada uno de los escenarios mencionados anteriormente se evaluaron las 3 bases de datos definidas y se usó una validación cruzada de 10 particiones.

3.3.1. Clasificación con máquinas de soporte vectorial a partir de características articulatorias

En esta primera etapa se pretende replicar los experimentos realizados en el estado del arte a partir de un esquema de aprendizaje de máquina tradicional. Inicialmente se toma una base de datos a la cual se le extrae las transiciones tanto onset como offset de cada hablante. Posteriormente se realiza una caracterización del sistema articulatorio a partir de los primeros 12 MFCCs, su primera y segunda derivada, y la energía de cada transición distribuida en 22 bandas de Bark. Finalmente, se calculan medidas estadísticas como la media, la desviación estándar, la asimetría y la kurtosis, y se construye un vector de características por cada grabación.

Para la etapa de clasificación se utilizó una máquina de soporte vectorial con un kernel Gaussiano, el cual es un algoritmo de clasificación que permite discriminar entre dos o mas clases a partir de un hiperplano de separación. Los parámetros del clasificador se optimizaron a través de una búsqueda por cuadrícula, con variaciones de $C \in \{0{,}001, 0{,}01, \cdots, 1000\}$ que controla la compensación entre el tamaño del margen y la penalización de los puntos ubicados en el otro lado del margen de decisión y variaciones de $\gamma \in \{0{,}0001, 0{,}001, \cdots, 100\}$ que corresponde al ancho de banda del kernel Gaussiano.

3.3.2. Entrenamiento y clasificación de CNNs con datos monolingües

La principal idea en esta etapa es construir un modelo robusto para cada base de datos que permita discriminar pacientes con EP de personas sanas.

Inicialmente se calcularon las transiciones de cada grabación de voz, con las cuales se construyeron espectrogramas a partir de la STFT, creando una imagen por transición, lo que genera la cantidad suficiente de datos para entrenar una red neuronal profunda. Luego de tener el total de espectrogramas de cada base de datos, se implementó una CNN con topología ResNet20 con 9 bloques residuales y 3 bloques principales, con 16, 32, y 64 mapas de características, respectivamente, un mayor detalle es mostrado en la Tabla 5. Para dar mayor claridad a los parámetros mostrados en la tabla, se toman las siguientes capas de ejemplo, para Conv(16 x 32 x 3, 2), 16 corresponde a los canales de entrada, 32 a los canales de salida, 3 al tamaño del kernel y 2 al stride, es decir, al paso del kernel. Avg Pool(11) realiza un pooling a partir de la media aritmética de los datos con un tamaño de kernel de 11x11. Para Lineal(64,2), 64 es el número de neuronas de entrada y 2 las neuronas de salida.

También se implementaron diferentes topologías de CNN basadas en LeNet [37] variando el número y tamaño de las capas convolucionales y lineales. La arquitectura que presentó mejor desempeño es mostrada en la Tabla 6, donde las capas convolucionales y lineales son equivalentes a las explicadas en la arquitectura ResNet. Max Pool(2) hace referencia a un Max pooling con un kernel de 2x2 y dropout al método de regularización para desactivar neuronas con cierta probabilidad. Para el entrenamiento de ambas redes se usó el método de Back-Propagation a partir del gradiente descendente usando como función de pérdida la entropía cruzada aplicando regularización L^2.

Tabla 5. Arquitectura ResNet20. Conv: Convolución, Avg Pool: Avg Pooling.

Etapa	**Tipo de capa**	**Tamaño de salida**
Entrada	Conv (1x16x3,1)	16x80x41
Bloque 1	Conv (16x16x3,1)	16x80x41
	Conv (16x16x3,1)	
	Conv (16x16x3,1)	
	Conv (16x16x3,1)	
	Conv (16x16x3,1)	
	Conv (16x16x3,1)	
Bloque 2	Conv (16x32x3,2)	32x40x21
	Conv (32x32x3,2)	
	Conv (32x32x3,2)	
	Conv (32x32x3,2)	
	Conv (32x32x3,2)	
	Conv (32x32x3,2)	
Bloque 3	Conv (32x64x3,2)	64x20x11
	Conv (64x64x3,2)	
	Conv (64x64x3,2)	
	Conv (64x64x3,2)	
	Conv (64x64x3,2)	
	Conv (64x64x3,2)	
Reducción	Avg Pool (11)	1x1x64
Salida	Lineal (64,2)	1x1x2

Tabla 6. Arquitectura CNN basadas en LeNet con mejor desempeño. Conv: Convolución, Max Pool: Max pooling.

Tipo de capa	**Tamaño de salida**
Conv (1x4x3,1) + dropout	4x80x41
Max Pool (2,2)	4x40x20
Conv (4x8x3,1) + dropout	8x40x20
Max Pool (2,2)	8x20x10
Conv (8x16x3,1) + dropout	16x20x10
Max Pool (2,2)	16x10x5
Conv (16x32x3,1) + dropout	32x10x5
Max Pool (2,2)	32x5x2
Lineal (320,128) + dropout	1x1x128
Lineal (128,64) + dropout	1x1x64
Lineal (64,2)	1x1x2

3.3.3. Clasificación a partir de aprendizaje por transferencia entre idiomas

En esta etapa, se implementó la clasificación de pacientes con EP vs. Controles sanos en modelos de CNNs entrenadas a partir de aprendizaje por transferencia. Se toma un modelo base, es decir, un modelo en un idioma específico y los parámetros de este sirven de base para re-entrenar la red con alguno de los dos idiomas restantes y no de forma aleatoria como se realiza comúnmente.

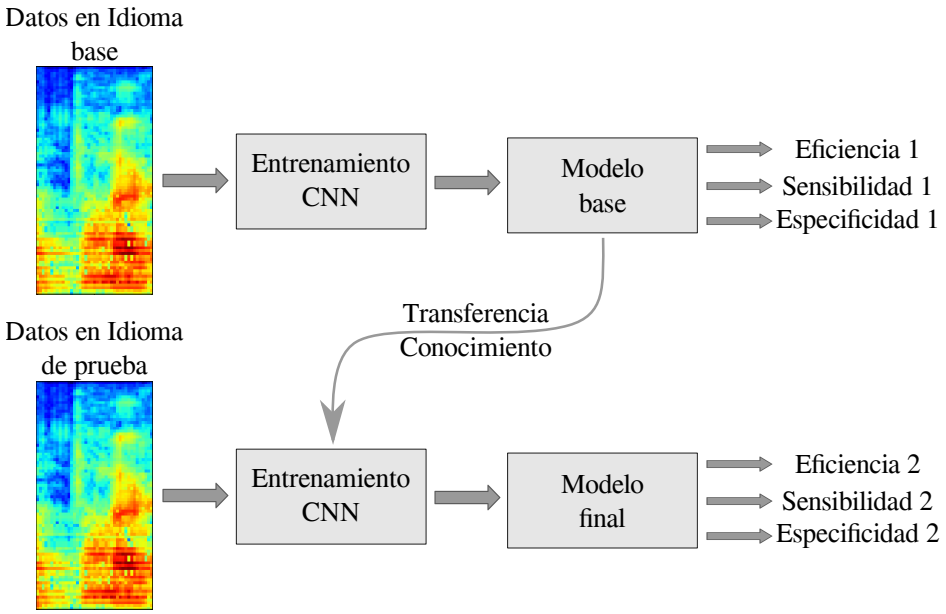


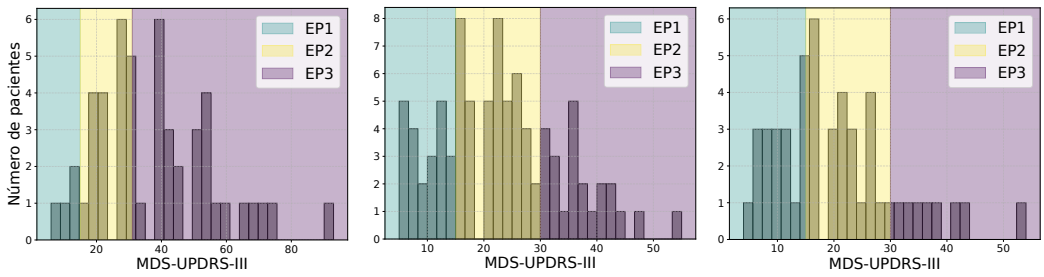
Figura 15. Transferencia de conocimiento en modelos de CNNs.

La Figura 15 resume la metodología desarrollada en esta etapa, la cual es el enfoque principal de este trabajo. En la parte superior de la figura se puede observar un diagrama de bloques de lo explicado en la sección 3.3.2, donde a partir de un espectrograma se entrena un modelo monolingüe y se evalúa este a partir de diferentes medidas de desempeño. En esta sección se añade un paso adicional, que hace referencia a la transferencia de conocimiento, un ejemplo para mayor claridad basándonos en esta figura sería que el modelo base corresponda al idioma Español. Posteriormente los parámetros de este modelo (parte superior) nos sirven para entrenar otra CNN (parte inferior) en alguno de los otros dos idiomas (Alemán o Checo). Finalmente lo que se espera es que los resultados de este experimento superen las medidas obtenidas en la clasificación de la CNN monolingüe.

3.3.4. Clasificación multiclase para monitorear el estado de severidad de los pacientes

Finalmente, se implementó un clasificador multiclase para predecir el estado de severidad de los pacientes con EP, para esto se dividió cada base de datos en cuatro grupos: (1) controles sanos; (2) pacientes con EP con puntuaciones MDS-UPDRS-III inferiores a 16 (etapa inicial – EP1); (3) pacientes con EP con puntuaciones MDS-UPDRS-III entre 16 y 30 (etapa intermedia – EP2); y (4) pacientes con EP con puntuaciones MDS-UPDRS-III superiores a 30 (etapa avanzada – EP3). En la Figura 16 se puede observar los histogramas del estado neurológico de los pacientes con EP según la escala MDS-UPDRS-III para cada idioma.

Para realizar este experimento se tomaron los mejores modelos de cada idioma implementados en la etapa de aprendizaje por transferencia. Para la predicción de los estados de severidad de los pacientes, se congelaron los parámetros de las capas convolucionales, es decir, la etapa de caracterización de la red, y solo se re-entrenaron las capas lineales con los espectrogramas correspondientes a su idioma. Como se puede observar en la Figura 16, existe un desbalance entre las clases, por lo tanto fue necesario usar "pesos" en la función de costo para cada grupo, con base en el porcentaje de muestras de cada clase, y así corregir el desbalance de clases en la etapa de entrenamiento. Finamente los resultados son reportados por idiomas a partir de una matriz de confusión, eficiencia y el coeficiente Kappa de Cohen.



(a) Español (b) Alemán (c) Checo

Figura 16. Histogramas del estado neurológico de los pacientes según su MDS-UPDRS-III. Pacientes en etapa inicial (verde), pacientes en etapa intermedia (amarillo) y pacientes en etapa avanzada (morado).

4. Resultados

4.1. Resultados monolingües

4.1.1. Máquinas de soporte vectorial

En la Tabla 7 es posible observar los resultados de la clasificación de pacientes con EP vs. Controles sanos usando características de articulación (MFCCs y energía en las bandas de Bark). Para la clasificación se entrenó y se evaluó una SVM por idioma, optimizando los parámetros C y γ para cada base de datos.

Se puede ver que los resultados son muy similares para las tres bases de datos en las 3 medidas de desempeño (eficiencia, sensibilidad y especificidad), sobresaliendo un poco el rendimiento del modelo entrenado con la base de datos Alemana con un promedio de 70.0 % en sus métricas.

Tabla 7. Clasificación de pacientes con EP vs. Controles sanos usando una SVM para diferentes idiomas. Efic: Eficiencia, Sens: Sensibilidad, Espec: Especificidad, μ: media, σ: desviación estándar.

Idioma	C	γ	Efic. ($\mu \pm \sigma$)	Sens. ($\mu \pm \sigma$)	Espec. ($\mu \pm \sigma$)
Español	1	0.0001	67.3 ± 2.4	65.2 ± 3.0	69.4 ± 3.1
Alemán	10	0.0001	69.9 ± 2.1	69.6 ± 2.2	70.2 ± 3.5
Checo	10	0.0001	68.3 ± 3.6	68.4 ± 4.3	68.1 ± 5.2

La idea principal de realizar este experimento es tener una referencia de los resultados obtenidos usando métodos de aprendizaje de máquina y observar como mejoran las medidas de desempeño al implementar aprendizaje profundo y por último el método de aprendizaje por transferencia.

4.1.2. CNN monolingües

Para la construcción de los modelos monolingües se implementaron 2 arquitecturas diferentes, una topología ResNet y una topología clásica.

La Tabla 8 muestra los resultados de la topología ResNet para cada base de datos. En esta tabla se puede observar que los resultados obtenidos en las bases de datos de Español y Alemán son bastante similares en su eficiencia pero sus principales variaciones se presentan en la sensibilidad y especificidad, lo que da a entender que el modelo Español permite discriminar de una mejor manera las personas sanas (Especificidad=84.0 %) y tiene falencias para clasificar pacientes con EP. Mientras que para el modelo Alemán su mejor resultado se presenta clasificando pacientes con EP

(Sensibilidad=74.8 %). Por último el modelo checo es acertado clasificando pacientes con EP (Sensibilidad=90.0 %), pero tiene bastantes dificultades para identificar los controles sanos.

Tabla 8. Clasificación de pacientes con EP vs. Controles sanos a partir de CNNs monolingües con topología ResNet. η: Tasa de aprendizaje, L^2: Regularización L^2, Efic: Eficiencia, Sens: Sensibilidad, Espec: Especificidad, μ: media, σ: desviación estándar.

Idioma	η	L^2	Efic. ($\mu \pm \sigma$)	Sens. ($\mu \pm \sigma$)	Espec. ($\mu \pm \sigma$)
Español	0.0001	0.0001	71.0 ± 11.0	58.0 ± 17.5	84.0 ± 15.8
Alemán	0.0005	0.0005	70.9 ± 9.9	74.8 ± 22.1	66.9 ± 15.9
Checo	0.0001	0.005	61.9 ± 12.0	90.0 ± 14.1	33.5 ± 29.1

Por otro lado, la Tabla 9 muestra los resultados obtenidos al implementar la CNN con la arquitectura clásica detallada en la Tabla 6. En estos resultados podemos observar que en general son similares a los obtenidos con la topología ResNet, con la diferencia que el modelo Español está más balanceado entre su especificidad-sensibilidad. El modelo Checo mejora su desempeño general, y por último el modelo Alemán redujo su eficiencia y su sensibilidad, pero incremento su especificidad.

Tabla 9. Clasificación de pacientes con EP vs. Controles sanos a partir de CNNs monolingües con arquitectura clásica η: Tasa de aprendizaje, Drop: Probabilidad de dropout, L^2: Regularización L^2, Efic: Eficiencia, Sens: Sensibilidad, Espec: Especificidad, μ: media, σ: desviación estándar.

Idioma	η	Drop.	L^2	Efic. ($\mu \pm \sigma$)	Sens. ($\mu \pm \sigma$)	Espec. ($\mu \pm \sigma$)
Español	0.005	0.3	0.0005	71.0 ± 15.9	74.0 ± 25.0	68.0 ± 28.6
Alemán	0.006	0.4	0.0005	63.1 ± 11.7	43.1 ± 38.0	83.1 ± 17.7
Checo	0.005	0.1	0.001	68.5 ± 14.1	94.0 ± 13.5	42.0 ± 33.2

Esto da a entender que ambas topologías tienen comportamientos similares para la clasificación de pacientes con EP y controles sanos, pero se decide usar como modelo base la arquitectura clásica, debido a que en mayoría sus resultados son un poco superiores. Además el costo computacional de esta arquitectura es mucho más bajo comparado con la cantidad de parámetros implementados en la topología ResNet. Quizás esta arquitectura tendría un mejor desempeño si las bases de datos tuvieran una mayor cantidad de imágenes (espectrogramas) de pacientes con EP, como se ha implementado en el estado del arte con bases de datos superiores a 1 millón de imágenes para la clasificación de animales.

4.2. Aprendizaje por transferencia entre idiomas

4.2.1. Español

Con el fin de mejorar los resultados obtenidos en los modelos monolingües se implementó la transferencia de conocimiento entre idiomas, en esta subsección nos enfocaremos en la base de datos de Español, por lo tanto se toma de modelo base los idiomas Checo y Alemán. Los resultados son reportados en la Tabla 10. Para más claridad se repitieron las medidas de desempeño del modelo monolingüe de Español, para realizar la comparación con mayor facilidad.

En esta tabla podemos observar que el mejor resultado se obtuvo cuando se tomaron los parámetros del modelo Checo y se re-entrenó con los datos en Español (Checo–Español), mejorando su eficiencia en 1 %, su especificidad en 10 % pero reduciendo su sensibilidad en 7 % con respecto al modelo Español. Tal vez los resultados obtenidos no son los mejores ya que el modelo Español obtuvo el mejor desempeño en el entrenamiento monolingüe y al momento de implementar la transferencia de conocimiento con un modelo de menor rendimiento, este no incrementa su desempeño.

Tabla 10. Clasificación de pacientes con EP vs. Controles sanos usando aprendizaje por transferencia para la base de datos de Español. η: Tasa de aprendizaje, Drop: Probabilidad de dropout, L^2: Regularización L^2, Efic: Eficiencia, Sens: Sensibilidad, Espec: Especificidad, μ: media, σ: desviación estándar.

Idioma	η	Drop	L^2	Efic. ($\mu \pm \sigma$)	Sens. ($\mu \pm \sigma$)	Espec. ($\mu \pm \sigma$)
Español	0.005	0.3	0.0005	71.0 ± 15.9	74.0 ± 25.0	68.0 ± 28.6
Checo–Español	0.005	0.3	0.0005	**72.0 ± 13.1**	67.0 ± 11.6	78.0 ± 23.9
Alemán–Español	0.005	0.3	0.0005	70.0 ± 12.5	62.0 ± 19.9	78.0 ± 29.0

Para una mejor interpretación de los resultados mostrados en la tabla anterior, la Figura 17 contiene la curva ROC de cada uno de los modelos. En esta gráfica se puede observar que el mejor desempeño se encuentra en el modelo Checo-Español con un área bajo la curva (del inglés Area Under the Curve, AUC) de 0.84 seguido por el modelo monolingüe con un AUC de 0.82. Por último la Figura 18 muestra los histogramas y la correspondiente distribución de densidad de probabilidad para el mejor modelo obtenido en la clasificación de pacientes y controles sanos en Español (Checo-Español). En esta figura se puede observar que se tiene un mayor error cuando se clasifica pacientes con EP, lo que equivale a los bins oscuros que se observan en el intervalo del umbral de decisión de 0 a 0.5.

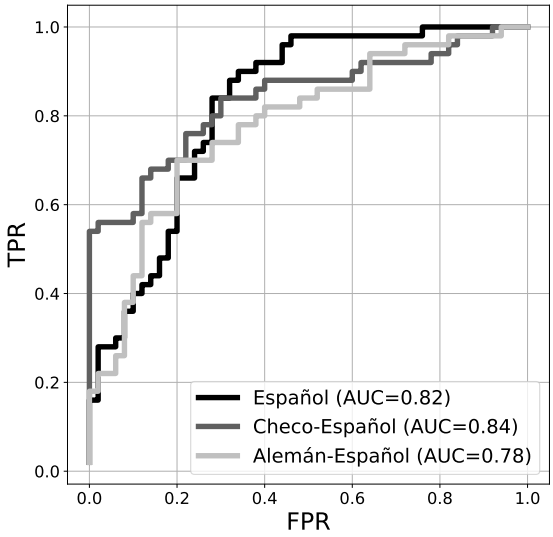


Figura 17. Curva ROC para Español.

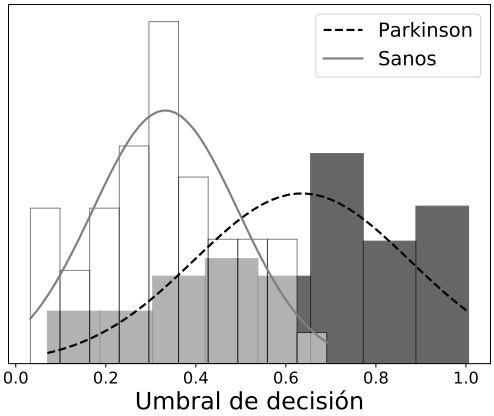


Figura 18. Histogramas y su correspondiente distribución de densidad de probabilidad para el modelo Checo-Español.

4.2.2. Alemán

Pasando a la clasificación de pacientes con EP vs. Controles sanos en la base de datos Alemana, se toma como modelos base los idiomas de Español y Checo. En la tabla Tabla 11 se resumen las medidas de desempeño al utilizar el método de aprendizaje por transferencia entre idiomas. En este experi-

mento se puede comprobar la mejora sustancial que logra la transferencia de conocimiento en los dos idiomas (Checo-Alemán y Español-Alemán). En promedio su eficiencia se incremento en un 14 % y su sensibilidad se duplico, a pesar de una reducción en su especificidad, lo cual logra generar modelos robustos que tienen una mayor capacidad para la discriminación de pacientes con EP.

Tabla 11. Clasificación de pacientes con EP vs. Controles sanos usando aprendizaje por transferencia para la base de datos Alemana. η: Tasa de aprendizaje, Drop: Probabilidad de dropout, L^2: Regularización L^2, Efic: Eficiencia, Sens: Sensibilidad, Espec: Especificidad, μ: media, σ: desviación estándar.

Idioma	η	Drop	L^2	Efic. ($\mu \pm \sigma$)	Sens. ($\mu \pm \sigma$)	Espec. ($\mu \pm \sigma$)
Alemán	0.006	0.4	0.0005	63.1 ± 11.7	43.1 ± 38.0	83.1 ± 17.7
Checo–Alemán	0.006	0.4	0.0005	76.7 ± 7.9	87.5 ± 11.0	66.0 ± 15.6
Español–Alemán	0.006	0.4	0.0005	**77.3 ± 11.3**	86.2 ± 13.8	68.3 ± 14.3

La Figura 19 muestra las curvas ROC de cada uno de los modelos, donde se puede visualizar el incremento en la sensibilidad del sistema monolingüe y se corrobora la superioridad de los modelos Checo-Alemán y Español-Alemán con un AUC de 0.79 y 0.82 respectivamente. Finalmente se construyó la distribución de densidad de probabilidad del mejor modelo (Español-Alemán) mostrada en la Figura 20, donde se puede ver que el mayor error de clasificación se presentó al discriminar controles sanos como pacientes con EP (baja especificidad) y que el error que comete la CNN al confundir pacientes con EP como controles sanos es mínimo (alta sensibilidad).

4.2.3. Checo

Por último para la clasificación de la base de datos Checa, se tomaron de base los modelos monolingües en Alemán y Español, con el fin de realizar aprendizaje por transferencia y mejorar el modelo Checo para discriminar pacientes con EP de controles sanos. La Tabla 12 muestra los resultados de este experimento y se concluye que la transferencia de conocimiento nuevamente mejora la robustez del sistema para la clasificación de la clase de interés. En este caso el mejor desempeño se obtuvo cuando se tomó de base el modelo Español y se re-entrenó con el idioma Checo (Español–Checo), mejorando su eficiencia en un 4 % y equilibrando las métricas de sensibilidad y especificidad, ya que el modelo monolingüe Checo tenia una alta sensibilidad pero muy baja especificidad.

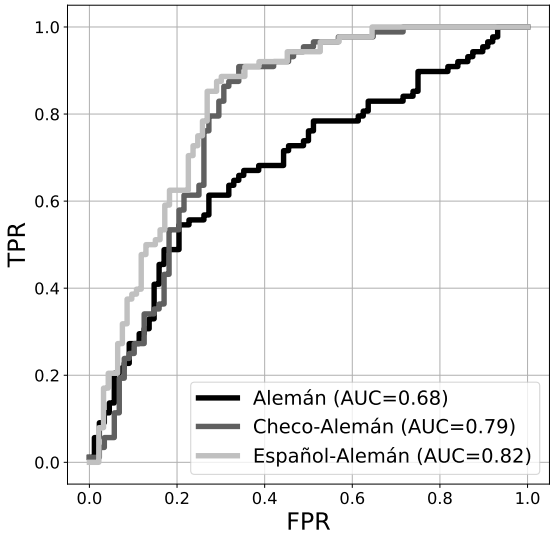


Figura 19. Curva ROC para Alemán.

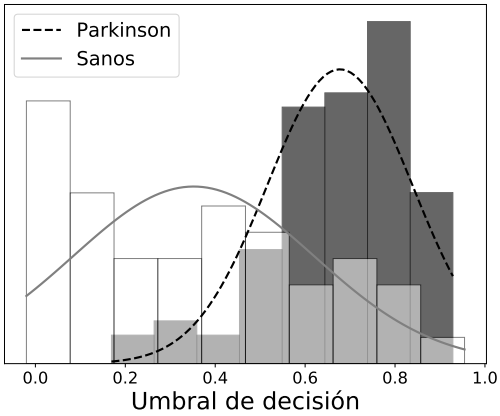


Figura 20. Histogramas y su correspondiente distribución de densidad de probabilidad para el modelo Español-Alemán.

En la Figura 21 se resumen los resultados reportados en la tabla anterior y se observa gráficamente la mejora que realiza el modelo Español sobre el modelo Checo (Español–Checo), pasando de un AUC de 0.76 del modelo monolingüe Checo a un AUC de 0.83. Finalmente para este último modelo se construyó la distribución de densidad de probabilidad con sus respectivos histogramas mostrados en la Figura 22, en donde es posible observar que

Tabla 12. Clasificación de pacientes con EP vs. Controles sanos usando aprendizaje por transferencia para la base de datos Checa. η: Tasa de aprendizaje, Drop: Probabilidad de dropout, L^2: Regularización L^2, Efic: Eficiencia, Sens: Sensibilidad, Espec: Especificidad, μ: media, σ: desviación estándar.

Idioma	η	Drop	L^2	Efic. ($\mu \pm \sigma$)	Sens. ($\mu \pm \sigma$)	Espec. ($\mu \pm \sigma$)
Checo	0.005	0.1	0.001	68.5 ± 14.1	94.0 ± 13.5	42.0 ± 33.2
Alemán–Checo	0.005	0.1	0.001	70.7 ± 14.5	80.0 ± 16.3	62.5 ± 26.3
Español–Checo	0.005	0.1	0.001	**72.6 ± 13.9**	82.0 ± 14.8	62.0 ± 28.9

el error de clasificación en ambas clases es más equilibrado que en las dos distribuciones anteriores, siendo un poco mayor la confusión de controles sanos como pacientes con EP, lo que corresponde a tener una especificidad menor.

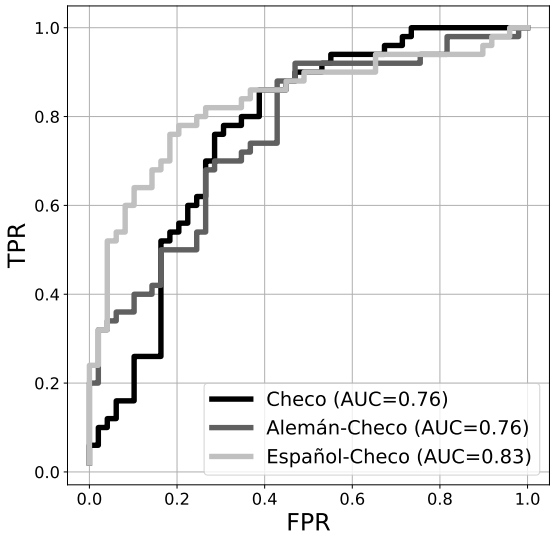


Figura 21. Curva ROC para Checo.

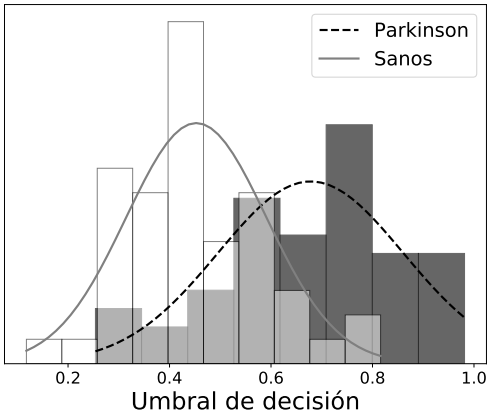


Figura 22. Histogramas y su correspondiente distribución de densidad de probabilidad para el modelo Español-Checo.

4.3. Evaluación del estado de severidad de los pacientes

Por último para predecir el estado de severidad de los pacientes con EP se usaron los modelos con mejor desempeño en la etapa de aprendizaje por transferencia, es decir, para Español se tomó el modelo Checo-Español, para Alemán el modelo Español-Alemán y para Checo el modelo Español-Checo. Los resultados son reportados en la Tabla 13, donde se muestran las matrices de confusión, la eficiencia y el coeficiente Kappa de Cohen para cada idioma.

Tabla 13. Matrices de confusión con los resultados de la clasificación de controles sanos y pacientes con EP en diferentes etapas de la enfermedad para diferentes idiomas. CS: Controles Sanos, EP1: Pacientes con puntuaciones MDS-UPDRS-III entre 0 y 15. EP2: Pacientes con puntuaciones MDS-UPDRS-III entre 16 y 30. EP3: Pacientes con puntuaciones MDS-UPDRS-III por encima de 30, Efic: Eficiencia, κ: Coeficiente kappa de Cohen. Las matrices de confusión están expresadas en porcentajes (%).

	Español				Alemán				Checo			
	Efic = 60.0		$\kappa = 0.38$		Efic = 50.6		$\kappa = 0.30$		Efic = 41.4		$\kappa = 0.13$	
	CS	EP1	EP2	EP3	CS	EP1	EP2	EP3	CS	EP1	EP2	EP3
CS	**72.0**	4.0	0.0	24.0	**51.1**	5.7	35.2	8.0	**57.1**	18.4	8.2	16.3
EP1	20.0	**20.0**	0.0	60.0	7.4	**14.8**	66.7	11.1	63.1	**21.1**	0.0	15.8
EP2	22.2	11.1	**5.6**	61.1	10.0	5.0	**80.0**	5.0	34.8	8.7	**21.7**	34.8
EP3	14.8	3.7	0.0	**81.5**	14.3	9.5	38.1	**38.1**	37.5	12.5	0.0	**50.0**

Los resultados en la tabla anterior, indican que el modelo Español sigue siendo el de mejor rendimiento, con una mayor eficiencia y un mejor coefi-

ciente κ, este modelo discrimina mejor los controles (72.0 %) y los pacientes en una etapa avanzada de la enfermedad (81.5 %), mientras que los otros dos estados de la enfermedad son confundidos en su mayoría como estados avanzados. Para el modelo Alemán se tiene un menor desempeño, pero con la particularidad de que la mayoría de los estados de severidad son confundidos con la etapa intermedia, incluyendo gran cantidad de los controles, lo que indica que no existe diferencia significativa para discriminar los diferentes estados de los pacientes. Por último, el idioma Checo es el de menor desempeño, el cual tiende a clasificar los 3 estados de severidad de los pacientes como controles sanos. Una posible solución para mejorar el desempeño de estos modelos, es incrementar el tamaño de las bases de datos, con el fin de que el sistema encuentre mejores características que permitan diferenciar fácilmente los estados de severidad de los pacientes.

5. Conclusiones

En este trabajo se implementó la técnica de aprendizaje por transferencia en CNNs a partir de tres idiomas diferentes para la clasificación de pacientes con EP y controles sanos. Inicialmente se crearon modelos monolingües a partir de espectrogramas extraídos de las transiciones onset y offset de señales de voz, luego se aplicó la transferencia de conocimiento entre idiomas para observar si es posible mejorar el rendimiento del clasificador teniendo como base un modelo en un idioma diferente. De acuerdo con los resultados reportados, se pudo comprobar que el aprendizaje por transferencia puede mejorar las medidas de desempeño de los modelos monolingües, con incrementos de hasta un 14 % en la eficiencia de éstos, indicando que los parámetros de una CNN en un idioma determinado son una buena base para re-entrenar CNNs en otros idiomas y a su vez generar modelos robustos que permitan la discriminación de pacientes con EP y controles sanos. Además se comprobó que es posible evaluar el estado de severidad de los pacientes a partir de los modelos creados por la transferencia de aprendizaje, obteniendo resultados de hasta un 60 % de eficiencia y un coeficiente kappa de 0.38.

El método de transferencia de conocimiento en otros idiomas obtiene buenos resultados siempre y cuando el modelo base sea bastante robusto, es decir, tenga un buen desempeño con sus datos de entrenamiento y prueba. Por tal motivo no fue posible mejorar el rendimiento del modelo Español, el cual tiene el mejor desempeño. Sin embargo, sirvió como base para mejorar sustancialmente el rendimiento de los otros modelos.

En general los resultados obtenidos para los diferentes experimentos, nos permiten concluir que el aprendizaje profundo o deep learning sobresale de estrategias de aprendizaje tradicionales como aquellas basadas en SVMs, teniendo en cuenta que el desempeño de la red neuronal depende de que sus datos de entrada sean lo suficientemente buenos, se utilice la arquitectura adecuada, y se implementen diferentes medidas de regularización evitando que la red tienda a sufrir un sobre-ajuste.

Como trabajo futuro se pretende crear modelos base más robustos, una de las formas es incrementando el número de datos de entrenamiento combinando 2 de las 3 bases de datos y realizando transferencia de conocimiento al idioma restante. También se pretende implementar un algoritmo de optimización bayesiana con el fin de obtener los parámetros óptimos para cada red y tener un mejor desempeño de esta. Adicionalmente, futuros experimentos incluirán aprendizaje por transferencia entre diferentes patologías, es decir, el modelo base construido para clasificar EP será usado para entrenar un clasificador para otro tipo de enfermedades neurodegenerativas como la enfermedad de Huntington.

6. Referencias

[1] O. Hornykiewicz, "Biochemical aspects of Parkinson's disease", *Neurology*, vol. 51, n.º 2 Suppl 2, S2-S9, 1998.

[2] J. A. Logemann, H. B. Fisher, B. Boshes y col., "Frequency and co-occurrence of vocal tract dysfunctions in the speech of a large sample of Parkinson patients", *Journal of Speech, Language, and Hearing Research*, vol. 43, n.º 1, págs. 47-57, 1978.

[3] C. Goetz, B. Tilley, S. Shaftman y col., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results", *Movement Disorders: Official Journal of the Movement Disorder Society*, vol. 23, n.º 15, págs. 2129-2170, 2008.

[4] B. T. Harel, M. S. Cannizzaro, H. Cohen y col., "Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment", *Journal of Neurolinguistics*, vol. 17, n.º 6, págs. 439-453, 2004.

[5] J. Rusz, R. Cmejla, H. Ruzickova y col., "Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease", *The Journal of the Acoustical Society of America*, vol. 129, n.º 1, págs. 350-367, 2011.

[6] C. Stewart, L. Winfield, A. Hunt y col., "Speech dysfunction in early Parkinson's disease", *Movement Disorders: Official Journal of the Movement Disorder Society*, vol. 10, n.º 5, págs. 562-565, 1995.

[7] Y. Yunusova, G. G. Weismer y M. J. Lindstrom, "Classifications of vocalic segments from articulatory kinematics: Healthy controls and speakers with dysarthria", *Journal of Speech, Language, and Hearing Research*, 2011.

[8] J. C. Vásquez-Correa, J. R. Orozco-Arroyave y E. Nöth, "Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease.", en *Interspeech*, 2017, págs. 314-318.

[9] J. R. Orozco-Arroyave, *Analysis of Speech of people with Parkinson's disease*. Logos Verlag Berlin GmbH, 2016, vol. 41.

[10] D. Montaña, Y. Campos-Roca y C. 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*, vol. 154, págs. 89-97, 2018.

[11] M. M. Hoehn y M. D. Yahr, "Parkinsonism: onset, progression, and mortality", *Neurology*, vol. 17, n.º 5, págs. 427-427, 1967.

[12] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave y col., "A Multitask Learning Approach to Assess the Dysarthria Severity in Patients with Parkinson's Disease.", en *Interspeech*, 2018, págs. 456-460.

[13] A. Naseer, M. Rani, S. Naz y col., "Refining Parkinson's neurological disorder identification through deep transfer learning", *Neural Computing and Applications*, págs. 1-16, 2019.

[14] D. Wang y T. F. Zheng, "Transfer learning for speech and language processing", en *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, IEEE, 2015, págs. 1225-1237.

[15] M. Oquab, L. Bottou, I. Laptev y col., "Learning and transferring mid-level image representations using convolutional neural networks", en *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, págs. 1717-1724.

[16] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla y col., "NeuroSpeech: An open-source software for Parkinson's speech analysis", *Digital Signal Processing*, vol. 77, págs. 207-221, 2018.

[17] E. Zwicker y E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency", *The Journal of the Acoustical Society of America*, vol. 68, n.º 5, págs. 1523-1525, 1980.

[18] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, n.º 3, págs. 235-238, 1977.

[19] J. B. Allen y L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis", *Proceedings of the IEEE*, vol. 65, n.º 11, págs. 1558-1564, 1977.

[20] B. Logan y col., "Mel Frequency Cepstral Coefficients for Music Modeling.", en *ISMIR*, vol. 270, 2000, págs. 1-11.

[21] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT press, 2016.

[22] J. Schmidhuber, "Deep learning in neural networks: An overview", *Neural Networks*, vol. 61, págs. 85-117, 2015.

[23] C. M. Bishop y col., *Neural Networks for Pattern Recognition*. Oxford university press, 1995.

[24] R. Reed y R. J. MarksII, *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks*. Mit Press, 1999.

[25] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1994.

[26] N. Srivastava, G. Hinton, A. Krizhevsky y col., "Dropout: a simple way to prevent neural networks from overfitting", *The Journal of Machine Learning Research*, vol. 15, n.º 1, págs. 1929-1958, 2014.

[27] G. Zaccone, M. R. Karim y A. Menshawy, *Deep Learning with TensorFlow*. Packt Publishing Ltd, 2017.

[28] K. He, X. Zhang, S. Ren y col., "Identity mappings in deep residual networks", en *European Conference on Computer Vision*, Springer, 2016, págs. 630-645.

[29] K. He, X. Zhang, S. Ren y col., "Deep residual learning for image recognition", en *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, págs. 770-778.

[30] D. Sarkar, R. Bali y T. Ghosh, *Hands-On Transfer Learning with Python: Implement advanced deep learning and neural network models using TensorFlow and Keras*. Packt Publishing Ltd, 2018.

[31] S. J. Pan y Q. Yang, "A survey on transfer learning", *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, n.º 10, págs. 1345-1359, 2010.

[32] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation", 2011.

[33] M. L. McHugh, "Interrater reliability: the kappa statistic", *Biochemia medica: Biochemia medica*, vol. 22, n.º 3, págs. 276-282, 2012.

[34] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. V. Bonilla y col., "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease.", en *LREC*, 2014, págs. 342-347.

[35] T. Bocklet, S. Steidl, E. Nöth y col., "Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues.", en *Interspeech*, 2013, págs. 1149-1153.

[36] J. Rusz, "Detecting speech disorders in early Parkinson's disease by acoustic analysis", 2018.

[37] Y. LeCun, L. Bottou, Y. Bengio y col., "Gradient-based learning applied to document recognition", *Proceedings of the IEEE*, vol. 86, n.º 11, págs. 2278-2324, 1998.