Arquitectura de aprendizaje profundo usando CNNs y RNNs para la clasificación de la enfermedad de Parkinson y Huntington a partir de señales de voz.

Autor:

Diego Alexander López Santander

Trabajo de Grado para optar al título de:

Ingeniero Electrónico

Asesor:

Prof. Dr.-Ing. Juan Rafael Orozco Arroyave

Co-asesor:

MSc. Cristian David Ríos Urrego

Línea de investigación:

Análisis de patrones y reconocimiento de señales

Grupo de investigación GITA

Universidad de Antioquia

Facultad de Ingeniería

Departamento de Ingeniería Electrónica y de Telecomunicaciones

Ingeniería Electrónica

Medellín, Antioquia, Colombia

2022

Cita	López-Santander [1]
Referencia **Estilo IEEE (2020)**	[1] López-Santander, D.A. “Arquitectura de aprendizaje profundo usando CNNs y RNNs para la clasificación de la enfermedad de Parkinson y Huntington a partir de señales de voz”, Trabajo de grado, Ingeniería Electrónica Universidad de Antioquia, Medellín, Colombia, 2022.

Grupo de Investigación en Telecomunicaciones Aplicadas (GITA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano/Director: Jesús Francisco Vargas Bonilla.

Jefe departamento: Augusto Enrique Salazar Jiménez.

Agradecimientos

En primer lugar, quiero agradecer a mis padres Elba Leonor Santander y Edilberto Leonel López, quienes han realizado un esfuerzo extraordinario por brindarnos a mis hermanos y a mí todas las herramientas necesarias para salir adelante. A ellos debo mi formación como persona y no hay palabras que alcancen a describir cuanto valoro su sabiduría y afecto. También agradezco a mis hermanos John López y Leidy López, quienes además de ofrecerme apoyo incondicional, han sido siempre mi ejemplo a seguir. A pesar de estar lejos de casa, mi familia ha estado siempre a mi lado.

En la Universidad de Antioquia he conocido a grandes personas y agradezco a cada una de ellas por las enseñanzas que me ha dejado. Especialmente deseo expresar mi gratitud hacia Jeferson Gallo, con quien he compartido una gran parte de mi formación académica y quien me invitó en primer lugar a formar parte del grupo de investigación GITA, en donde he encontrado un ambiente de aprendizaje, apoyo y compañerismo. En general, agradezco a todo el grupo GITA por acogerme y darme una oportunidad para crecer profesionalmente.

Finalmente quiero agradecer a mi asesor Rafael Orozco y a mi co-asesor Cristian Rios por confiar en mí, por expresar siempre paciencia y compresión, por compartir todo su conocimiento y experiencia, y por acompañarme en cada paso, no solamente de la realización de este trabajo de grado, sino también de mi formación académica y personal en el camino de la investigación. Nada de esto sería posible sin ustedes. ¡Gracias!

Tabla de contenido

Lista de tablas

Lista de figuras

Resumen

Los desórdenes neurodegenerativos como las enfermedades de Parkinson o de Huntington afectan las funciones normales del cuerpo como el habla, el movimiento, el equilibrio, entre otros. Específicamente el deterioro del habla se produce por la pérdida del control de los músculos encargados de la producción del lenguaje oral, esta condición se denomina disartria. Teniendo en cuenta que la disartria está ligada con frecuencia a la progresión de estas enfermedades y que cada una provoca distintos tipos de disartria (disartria hipocinética e hipercinética para Parkinson y Huntington respectivamente), es posible desarrollar sistemas de evaluación automática a partir de señales de voz para apoyar a los profesionales de la salud en el diagnóstico y toma de decisiones para el tratamiento temprano de pacientes de enfermedades neurodegenerativas como el Parkinson y Huntington.

El enfoque propuesto en el presente trabajo consiste en desarrollar una arquitectura basada en redes neuronales convolucionales de una dimensión (CNNs) seguidas de redes neuronales recurrentes (RNN) para la clasificación del habla patológica, considerando que esta configuración es típicamente usada para el modelamiento de información secuencial, como es el caso de una señal de audio en el dominio del tiempo. Particularmente, en este trabajo se realizó la clasificación de la disartria hipocinética vs. habla sana, disartria hipercinética vs. habla sana y disartria hipocinética vs disartria hipercinética.

El modelo desarrollado fue entrenado y evaluado usando dos bases de datos con diferentes tareas de habla realizadas por hablantes nativos checos. Además, se comparó el rendimiento del modelo implementado con métodos clásicos, es decir, sin el uso de herramientas de aprendizaje profundo. En general los resultados alcanzados con la arquitectura de aprendizaje profundo propuesta no superaron los resultados obtenidos usando características clásicas de articulación y prosodia, sin embargo, se desarrolló un marco de trabajo sistemático con el potencial para evaluar y optimizar modelos de aprendizaje profundo.

Palabras clave: Disartria, Enfermedad de Parkinson, Enfermedad de Huntington, Habla, Aprendizaje profundo, Redes Neuronales Convolucionales, Redes Neuronales Recurrentes.

Abstract

Neurodegenerative disorders such as Parkinson's or Huntington's disease affect normal body functions such as speech, movement, balance, among others. Specifically, speech deterioration is caused by the loss of proper control of the muscles responsible for oral language production, this condition is called dysarthria. Considering that dysarthria is often linked to the progression of these diseases and that each causes different types of dysarthria (hypokinetic and hyperkinetic dysarthria for Parkinson's and Huntington's disease respectively), it is possible to develop automatic evaluation systems based on speech signals to support healthcare professionals in diagnosing and making decisions for early treatment of patients with neurodegenerative diseases such as Parkinson's and Huntington's disease.

The approach proposed in this work consists of developing an architecture based on one-dimensional convolutional neural networks (CNNs) followed by recurrent neural networks (RNNs) for pathological speech classification, considering that this configuration is typically used for modeling sequential information, such as an audio signal in the time domain. Particularly, the experiments performed in this work were: the classification of hypokinetic dysarthria vs. healthy speech, hyperkinetic dysarthria vs. healthy speech and hypokinetic dysarthria vs. hyperkinetic dysarthria.

The developed model was trained and evaluated using two databases with different speech tasks performed by Czech native speakers. In addition, the performance of the implemented model was compared with classical methods, i.e., without the use of deep learning tools. Overall, the results achieved with the proposed deep learning architecture did not outperform the results obtained using classical articulation and prosody features, however, the systematic framework developed throughout the research has the potential to evaluate and optimize other deep learning models.

Keywords: Dysarthria, Parkinson's Disease, Huntington's Disease, Speech, Deep Learning, Convolutional Neural Networks, Recurrent Neural Networks.

1 Introducción

Las enfermedades de Parkinson (PD, del inglés *Parkinson's Disease*) y Huntington (HD del inglés *Huntington's Disease*) son desórdenes neurodegenerativos que afectan gravemente la calidad de vida de los pacientes [1]. La PD se caracteriza por temblores en reposo, bradicinesia, es decir, lentitud en los movimientos, rigidez y congelamiento en la marcha [2]. Por otro lado, la HD produce movimientos involuntarios abruptos, irregulares y de corta duración denominados corea, especialmente en las manos y la cara [3]. Debido a que el habla es una tarea motora compleja que requiere la sincronización de diversos músculos cuyo funcionamiento correcto se ve afectado por ambas enfermedades, un síntoma común de ambas enfermedades es la disartria; un trastorno del habla caracterizado por la dificultad en el control del aparato fonador. PD y HD causan dos tipos distintos de disartria: disartria hipocinética en el caso de PD e hipercinética para HD. La disartria hipocinética se caracteriza por la monotonía de la voz, falta de fluidez, temblor en la voz y pronunciación imprecisa, entre otros síntomas [4]. Por otro lado, los síntomas de la disartria hipercinética incluyen interrupciones aleatorias de articulación, prosodia anormal e hipernasalidad [5].

Debido a que los trastornos del lenguaje oral son un síntoma común en PD y HD, el habla de un individuo se puede utilizar como indicador para el desarrollo de herramientas computarizadas para apoyar al diagnóstico y monitoreo de pacientes. Además, se tiene la ventaja de que las grabaciones de voz se pueden obtener con relativa facilidad, un costo bajo y sin procedimientos invasivos. Por otro lado, el habla de un paciente puede permitir reconocer la patología, considerando que la PD y HD tienen asociadas distintos tipos de disartria. Para la captura de señales de voz se suelen usar diversas tareas para evaluar diferentes parámetros de las patologías. Las tareas más comunes incluyen lectura en voz alta de textos, pronunciación de vocales sostenidas o moduladas, monólogos cortos y la repetición rápida de tareas diadococinéticas (DDK), es decir, palabras con combinaciones de consonantes plosivas y vocales [6].

Actualmente existe un gran interés en la aplicación de herramientas de aprendizaje profundo en diversas áreas por su gran potencial y versatilidad para la resolución de problemas de manera automática. Particularmente, para la clasificación de señales de habla patológica, el

aprendizaje profundo permite la definición de modelos que operan a partir de las señales originales (o con un mínimo preprocesamiento). De esta manera, se suprime la necesidad de extraer manualmente características definidas por profesionales. Sin embargo, entre las desventajas del aprendizaje profundo se encuentra la difícil interpretabilidad de los sistemas y la gran cantidad de recursos computacionales necesarios para el entrenamiento.

Dado el interés de la comunidad científica por utilizar el habla como un indicador de la presencia de enfermedades neurodegenerativas, existen varios estudios en la literatura donde se utilizan diversas metodologías para clasificar pacientes de HD, PD e individuos de control sanos (HC) e identificar las manifestaciones de la enfermedad en el habla. En Rusz et al. [7] los autores recomiendan una metodología estandarizada para el registro del habla de pacientes con disartria hipocinética o hipercinética asociadas a PD y HD respectivamente. En dicho estudio, se concluye que el grupo de pacientes con disartria hipercinética presenta una mayor cantidad de dimensiones del habla afectadas en comparación con el grupo con disartria hipocinética. En Novotny et al. [8] se investigó la relación de la hipernasalidad en el habla con PD y HD, encontrando que la nasalidad anómala no es característica de PD mientras que los pacientes de HD presentan con frecuencia hipernasalidad intermitente. Por otro lado, Hlavnička et al. [9] caracterizaron temblores de la voz para una gran variedad de trastornos neurológicos, incluyendo HD y PD. Se encontró que para 65% de los pacientes de HD y 20% de los pacientes de PD presentan temblor en la voz. Muchos estudios se basan en técnicas clásicas de extracción de características creadas a mano y el uso de clasificadores como máquinas de soporte vectorial (SVM), K vecinos más cercanos o bosques aleatorios [6]. En Moro-Velazques et al. [10] los autores presentan un modelo de extracción de diferentes características acústicas para entrenar un modelo de mezclas gaussianas (GMM) usando tres bases de datos. Se reportan tasas de acierto de entre 81% y 94% para la clasificación dependiendo de la base de datos y hasta 76% para la clasificación entre distintas bases de datos. En Solana-Lavalle et al. [11] se realiza la clasificación de pacientes de PD usando características clásicas para tareas de habla: Coeficientes Cepstrales en la escala de Mel (MFCC), medidas de fonación, coeficientes de wavelet. Se implementaron varios clasificadores obteniendo una tasa de acierto máxima de 94%. En Vásquez-Correa et al. [12],[13] se plantea el uso de redes neuronales convolucionales y aprendizaje por transferencia para la clasificación de PD y HC a partir de señales de voz en diferentes idiomas. Inicialmente logran tasas de acierto de 71% y 81% respectivamente

para PD y HC con redes neuronales entrenadas con cada base de datos de forma individual, posteriormente los resultados se logran mejorar a 83% y 86% utilizando técnicas de transferencia de aprendizaje entre idiomas y entre enfermedades. En Saeed-Mian [14] se evaluó el rendimiento de una arquitectura de redes neuronales convolucionales en una dimensión (CNN-1D) para la detección de PD a partir de señales de audio reportando una tasa de acierto de 92%. Arquitecturas con CNN-1D y redes neuronales recurrentes (RNN) se han propuesto recientemente para la identificación de PD a partir de señales de voz y para una clasificación multiclase del nivel de disartria de los pacientes, en Rios-Urrego et al. [15] se obtiene un resultado 89% en la exactitud de predicciones biclase (PD y HC) y de 60% la clasificación multiclase de los cuatro niveles de disartria de la escala m-FDA, utilizando arquitecturas de este tipo. En Mallela et al. [16] se propone una arquitectura de CNN seguida de una LSTM para la clasificación de PD y esclerosis lateral amiotrófica obteniendo una tasa de acierto de 88.5%.

En este trabajo se realizó una clasificación automática de pacientes con EP y HD, con HC usando una arquitectura de CNNs y RNNs, evaluando que tan efectivas resultan para la clasificación del tipo de disartria causada por cada trastorno, a partir de señales de voz. Cabe resaltar que las topologías de redes neuronales propuestas son frecuentemente utilizadas para tratar información secuencial, razón por la cual son una buena opción para tratar con señales de voz, las cuales se representan como un vector de amplitudes que varían en el tiempo. El modelo obtenido se entrenó específicamente usando dos bases de datos de hablantes nativos checos, sin embargo, este puede ser generalizado fácilmente para trabajar en la clasificación de las enfermedades PD y HD en otros idiomas. Los resultados obtenidos son comparados con modelos de referencia basados en características clásicas de articulación y prosodia. Se encontró que la arquitectura planteada no es capaz de superar el desempeño del modelo clásico de referencia.

2 Objetivos

2.1 Objetivo general

Evaluar el desempeño de modelos de clasificación basados en redes neuronales convolucionales (CNNs) y redes neuronales recurrentes (RNNs) para identificar la enfermedad de Huntington y Parkinson a partir de señales de voz.

2.2 Objetivos específicos

- Definir las características de la arquitectura que se usará para la clasificación automática de pacientes y controles a partir de diferentes topologías de la red neuronal considerando variables como el tamaño de las ventanas de audio para los filtros y diferentes técnicas de regularización.
- Implementar y evaluar el desempeño de una arquitectura de redes neuronales convolucionales seguida de redes neuronales recurrentes para la clasificación de la enfermedad de Parkinson y Huntington.
- Comparar, a partir de diferentes medidas de desempeño, los modelos de aprendizaje profundo desarrollados tomando como referencia técnicas tradicionales de aprendizaje automático.

3 Marco teórico

3.1 Dimensiones del habla

Son particularidades de la expresión oral que pueden ser medidas y estudiadas para caracterizar el habla de una persona, permitiendo detectar variaciones inusuales que pueden ser el resultado de desórdenes del lenguaje. Específicamente, en este trabajo se utilizan características de **prosodia y articulación** para identificar patrones de habla causados por las enfermedades de Huntington y Parkinson asociadas a la disartria hipercinética e hipocinética, respectivamente.

La **prosodia** es la variación del volumen, el tono, tiempo y ritmo para producir un habla natural [17]. Usando el *toolkit* DisVoice[1] se ralizó la extracción de 103 características de prosodia, las cuales incluyen la media, desviación estándar, asimetría, curtosis, valor máximo y mínimo de la duración de segmentos sonoros, duración de segmentos sordos, duración de las pausas, energía de segmentos sordos, energía de segmentos sonoros, el contorno de la frecuencia fundamental, entre otros, estas características están basadas en Najim Dehak [18].

La **articulación** comprende los cambios de posición, tensión y forma de los órganos, tejidos y extremidades que intervienen en la producción del habla [17]. Con ayuda del *toolkit* DisVoice se extraen 122 descriptores de articulación incluyendo 12 Coeficientes cepstrales en la escala de Mel (abreviados MFCC por sus siglas en inglés), la primera y segunda derivada de los MFCCs, la frecuencia del primer formante, la frecuencia del segundo formante y sus derivadas de primer y segundo orden, entre otros.

3.2 CNNs

Una red neuronal convolucional (CNN, del inglés *Convolutional Neural Network*) es un algoritmo de aprendizaje profundo que toma como entrada una representación matricial de datos, por ejemplo, una imagen y permite entrenar ciertos parámetros denominados pesos y sesgos (nombrados comúnmente en inglés: *weights and biases*) para identificar regiones u objetos dentro

[1] https://disvoice.readthedocs.io/en/latest/

de la matriz y diferenciar entre distintos datos [19]. El preprocesamiento necesario en una CNN es mucho menor comparado con otros algoritmos de clasificación debido a que los algoritmos modernos permiten entrenar automáticamente filtros o características que en otros algoritmos se deben diseñar a mano. Normalmente una CNN consta de tres etapas: La primera etapa consiste en realizar varias convoluciones en paralelo para obtener una matriz más representativa de las características críticas para realizar una buena predicción. La segunda etapa consiste en una función de reducción o *pooling*, la cual disminuye el tamaño de la matriz buscando extraer los parámetros más relevantes. Finalmente, la etapa de clasificación es una red totalmente conectada que da como resultado la predicción final del sistema.

3.2.1 Etapa de convolución

La operación de convolución realizada en esta capa busca extraer las características de alto nivel de la entrada. Normalmente la convolución se realiza en repetidas ocasiones. Las primeras capas de convolución identifican características de bajo nivel y cada capa de convolución extra se encarga de integrarlas para alcanzar características de alto nivel. De esta manera el sistema genera una comprensión completa de la entrada en diferentes escalas.

La operación de convolución se realiza entre la matriz de información I y la matriz *Kernel* K como se indica en la *ecuación 1*.

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \quad (1)$$

De forma gráfica la convolución se puede ver como una multiplicación de una región de la entrada con la matriz *Kernel*, mientras esta se desplaza tantas veces sean necesarias para cubrir la totalidad de la entrada como se muestra en la ***Figura 1.*** De esta forma se pueden modelar pequeñas piezas de información que son combinadas en capas posteriores de la red para comprender la entrada [19].

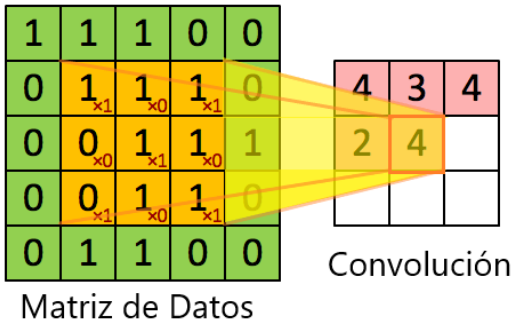

Figura 1. Representación gráfica de una capa de convolución

3.2.2 Etapa de reducción o pooling

La etapa de reducción o *pooling* se utiliza normalmente después de la etapa de convolución y permite reducir las dimensiones espaciales de la entrada mediante un submuestreo o resumen estadístico. Entre las ventajas generadas por el proceso de *pooling* está la reducción de los recursos computacionales requeridos para el funcionamiento del sistema, la extracción de las características más predominantes de la entrada y la eliminación del exceso de información que puede producir sobreajuste en la salida.

Existen varias clases de *pooling*, principalmente: *Max Pooling* y *average Pooling*. *Max pooling* da como resultado el valor máximo de la región contenida en el *Kernel*, mientras que *Average Pooling* devuelve el promedio de los valores cubiertos por el *Kernel* [19].

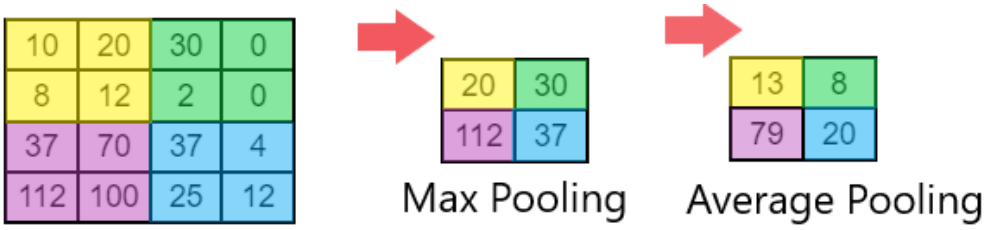

Figura 2. Etapa de reducción o *pooling*

3.2.3 Etapa clasificación

Finalmente, una capa neuronal completamente conectada es una forma común de interpretar las características de alto nivel identificadas por las capas convolucionales y de *pooling* para realizar una clasificación. Esta capa completamente conectada se denomina perceptrón multicapa y requiere una entrada de una dimensión, por lo tanto, la salida matricial de las capas convolucionales debe convertirse (*flatten*) en un vector columna.

La red se entrena mediante *backpropagation* en repetidas iteraciones para identificar patrones en la entrada y realizar una clasificación. El número de neuronas en la capa de clasificación es igual al número de clases que se debe predecir [19].

3.3 CNN-1D

Las redes neuronales convolucionales de una dimensión son frecuentemente utilizadas para modelar datos secuenciales, debido a que estos suelen ser representados como vectores de datos de una sola dimensión. Cada capa convolucional está compuesta por un número de canales y un *kernel* para cada uno de ellos. El *kernel* consiste en un filtro que recorre la entrada mediante la operación de correlación cruzada y cuyos parámetros son ajustados durante el proceso de entrenamiento. El tamaño del *kernel* determina fundamentalmente la longitud de la ventana temporal que es analizada por cada filtro convolucional, específicamente, para una señal con frecuencia de muestreo f_s y un *kernel* de tamaño n, el tamaño de la ventana esta dado por el cociente n/f_s. Para una capa convolucional de tamaño (N, C_{in}, L) la salida (N, C_{out}, L) se describe mediante la *ecuación 3*:

$$out(N_i, C_{out}) = bias\left(C_{out\ j}\right) + \sum_{k=0}^{C_{in}-1} weight\left(C_{out\ j}, k\right) * input(N_i, k) \qquad (3)$$

Donde $*$ denota la operación de correlación cruzada, N es el tamaño del lote, C es el número de canales y L es la longitud de las señales secuenciales [20].

Para este tipo de redes neuronales convolucionales la etapa de reducción o *pooling* corresponde a un submuestreo temporal que depende del tamaño del *kernel*. Para una señal de entrada de longitud L, con un *kernel* de longitud n y deslizamiento igual al tamaño del *kernel*; la salida tendrá una longitud igual al cociente entre L y n debido a que por cada n muestras de la señal, la salida se encuentra como el valor máximo o el valor promedio de las muestras cubiertas por el *kernel*, dependiendo si se usa la estrategia de *max pooling* o *average pooling* respectivamente.

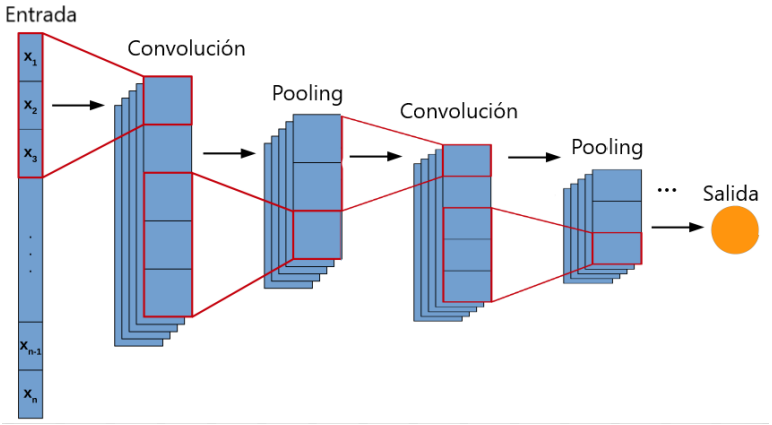

Figura 3. Red Neuronal Convolución 1D

Una ventaja de las CNN-1D es la gran diferencia en cuanto a la complejidad computacional de las convoluciones de una y dos dimensiones, es decir, una señal 2D con dimensiones $N \times N$ convolucionada con un *kernel* de tamaño $K \times K$ tendrá una complejidad computacional $\sim O(N^2K^2)$ mientras que en la correspondiente convolución 1D (con las mismas dimensiones, N y K) ésta es $\sim O(NK)$. Esto significa que en condiciones equivalentes la complejidad computacional de una CNN-1D es significativamente menor que la de una CNN-2D [21].

3.4 RNN-LSTM

Las redes neuronales recurrentes (RNN, del inglés *Recurrent Neural Network*) son arquitecturas donde una salida previa puede ser usada como la entrada de un paso posterior. De esta forma las RNNs son capaces de capturar el comportamiento temporal de una secuencia. Sin embargo, se ha demostrado que cuando se intenta analizar intervalos de tiempo muy amplios con

demasiadas etapas de recursión, el algoritmo de descenso gradiente presenta problemas de convergencia, resultando en dificultades para entrenar las RNNs [22].

Para mejorar el desempeño de las RNNs se desarrolla la arquitectura de redes neuronales recurrentes LSTM (del inglés Long-*Short Term Memory*), que introducen una celda de memoria y un mecanismo de compuertas para mejorar su capacidad de "recordar" largas secuencias temporales. El estado de la celda de memoria se actualiza en función de tres compuertas [23]:

- La compuerta f_t, controla cuanta información del anterior estado debe ser retenida por la LSTM.
- La compuerta i_t, controla cuanta información nueva debe ser agregada a la memoria de las LSTM
- La compuerta N_t, controla cuanta memoria se usará en el estado de salida para el paso actual.

En cada celda de una LSTM se debe determinar la información que debe obtenerse de la misma. La función logística sigmoide permite decidir qué parte de la salida debe ser extraída. La compuerta olvidar (f_t, *ecuación 4*) tiene como entrada h_{t-1} que es un vector con valores entre 0 a 1.

$$f_t = \sigma(W_F[h_{t-1}, X_t] + b_f) \quad (4)$$

Donde σ, W_F y b_f indican la función sigmoide, las matrices de pesos y el vector de sesgos de la compuerta, respectivamente. El cálculo del siguiente paso en una unidad LSTM se dan en las ecuaciones 4-6. En la *ecuación 5*, se almacena la información X_t de la entrada anterior y se actualiza el estado de la celda. La etapa de entrada incluye dos partes que son la capa sigmoide y la capa *tanh*. En la *ecuación 6* se usa función de activación tangente hiperbólica para forzar a los valores a un rango entre - 1 y 1. La nueva información adquirida, C_{t-1} se añade en el estado actual de la celda C_t como se muestra en la *ecuación 7* [24].

$$i_t = \sigma(W_i[h_{t-1}, X_t + b_i]) \quad (5)$$

$$N_t = tanh(W_n[h_{t-1}, X_t] + b_n) \quad (6)$$

$$C_t = C_{t-1}\, f_t + N_t i_t \quad (7)$$

El esquema de la ***Figura 4*** muestra tanto la naturaleza secuencial de un conjunto de unidades LSTM, como las 3 compuertas principales que rigen el funcionamiento de cada unidad LSTM.

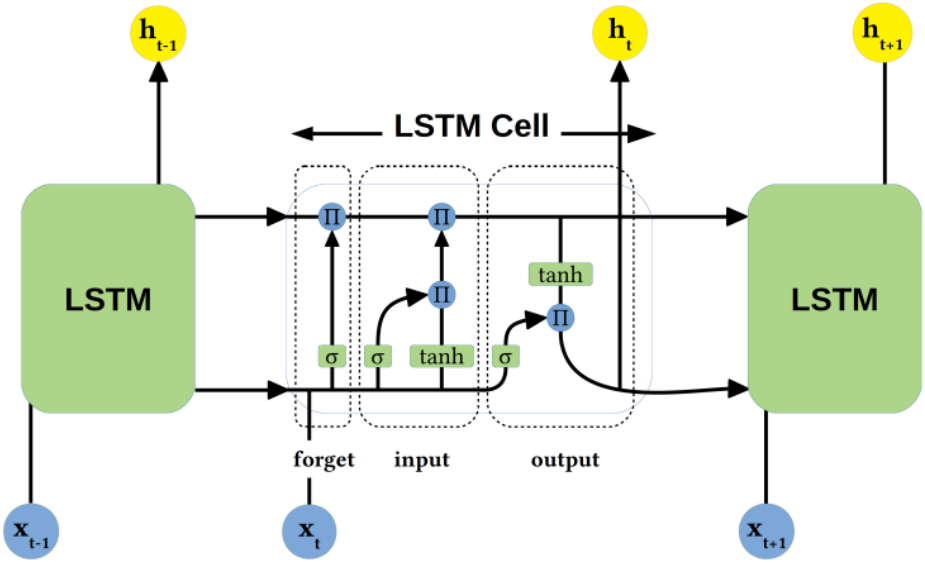

Figura 4. Arquitectura básica de una LSTM. Tomado de [23]

3.5. Regularización

La regularización es un proceso importante en el entrenamiento de redes neuronales profundas, debido a que permite limitar uno de los grandes problemas que estas pueden presentar: el sobre ajuste. Específicamente, la regularización busca generar modelos robustos y con capacidad de generalización ante la entrada de nuevos datos. En el presente trabajo se usan 3 tipos de regularización:

Regularización L2

También conocida como regresión de cresta, consiste en la adición de un término a la función de costo utilizada para el entrenamiento de la red. El término adicional consiste en una penalización que incrementa con la complejidad del modelo y es proporcional a la magnitud cuadrada de los pesos de la red. Matemáticamente, la función de costo modificada se expresa de la siguiente manera:

$$C'(w) = C(w) + \frac{1}{2}\lambda|w|^2 \qquad (2)$$

Dónde $C(w)$ representa la función de costo original, λ es un parámetro adicional que permite modificar el peso de la penalización y w son los pesos de la red [25].

Early Stopping

Durante el entrenamiento de una red neuronal se cuenta con un conjunto de entrenamiento y otro de prueba para comprobar que el modelo es capaz de aprender del primer conjunto y aplicar el nuevo conocimiento para realizar predicciones sobre el segundo. Comúnmente se espera que durante las primeras épocas de entrenamiento, la pérdida de ambos conjuntos se reduzca gradualmente. Posteriormente, cuando ocurre el sobreajuste, la perdida de prueba tiende a incrementar mientras que la perdida de entrenamiento continúa disminuyendo. El algoritmo de detección temprana (*early stopping*) consiste en guardar los parámetros del modelo cada vez que la pérdida de prueba mejore y detener forzosamente el entrenamiento después de un número determinado de épocas sin mejoría con el fin de obtener el mejor modelo antes de alcanzar un sobre ajuste.

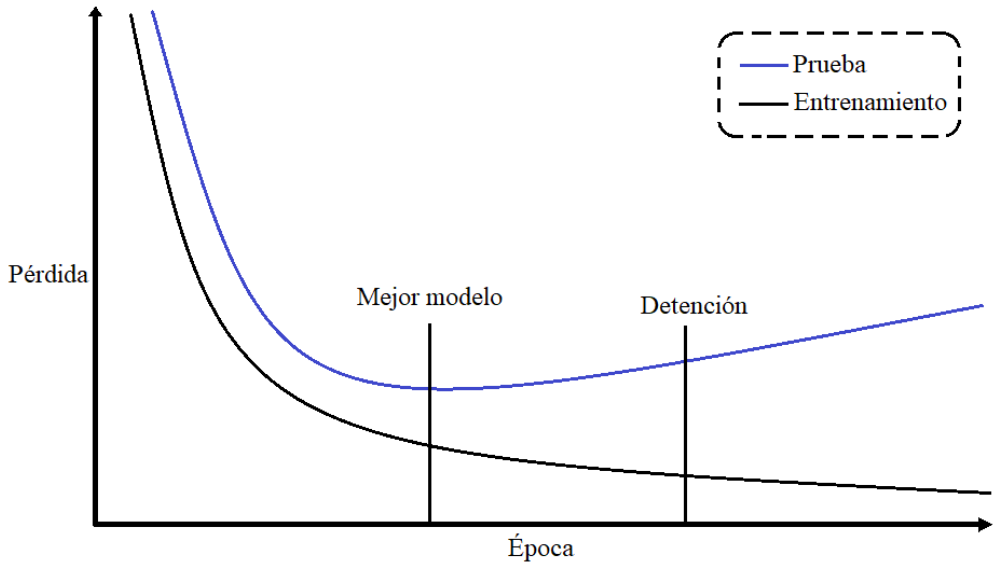

Figura 5. *Early Stopping*

Dropout

El *dropout* consiste en desactivar aleatoriamente un cierto número de conexiones entre neuronas, de modo que las neuronas desactivadas en cada iteración no participan en el proceso de entrenamiento (actualización de pesas) [26]. Este procedimiento reduce las dependencias entre neuronas vecinas y evita que se presente el sobre ajuste. Para cada conexión entre neuronas se define una variable aleatoria de Bernoulli, la cual determina cual es la probabilidad de que dicha conexión permanezca o sea desactivada.

3.6 Validación Cruzada

La validación cruzada es una técnica usada en el aprendizaje de máquina para evaluar de forma robusta el desempeño de un modelo en un experimento. Mediante la validación cruzada se garantiza que los resultados son independientes de la elección particular de conjuntos de prueba y entrenamiento debido a que, durante las iteraciones del procedimiento, cada elemento del conjunto de datos participa (en iteraciones diferentes) tanto del conjunto de prueba, como del conjunto de entrenamiento.

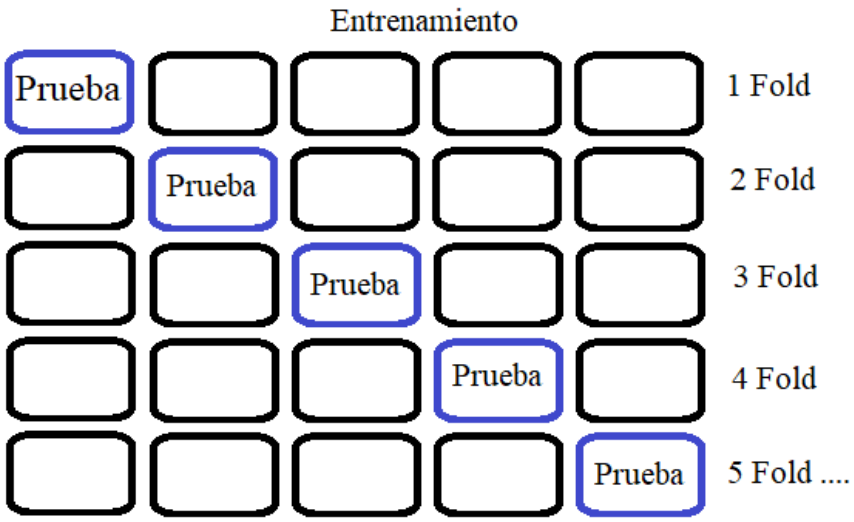

Figura 6. Esquema de validación cruzada

Como se muestra en la ***Figura 6***, durante la primera iteración del procedimiento de validación cruzada (denominado *Fold*), el conjunto de datos se divide en un conjunto de prueba independiente de tamaño igual al $(1/k)\%$ del total de datos, mientras que el resto se toman para el conjunto de entrenamiento. Las pruebas necesarias son realizadas en base en la partición de datos realizada y se repite el proceso k veces, realizando en cada *Fold* una elección diferente de los conjuntos de prueba y entrenamiento.

3.7 Máquinas de soporte vectorial

Las máquinas de soporte vectorial (SVM, del inglés *Support Vector Machine*) son algoritmos de clasificación que operan mediante el trazo de hiperplanos de separación. Para un espacio con n dimensiones, un hiperplano de separación consiste en un plano de dimensión n – 1 que idealmente contiene a cada lado las muestras correspondientes a una sola clase. Si existe dicho hiperplano que aísla perfectamente una clase de la otra, se dice que el conjunto de datos es linealmente separable.

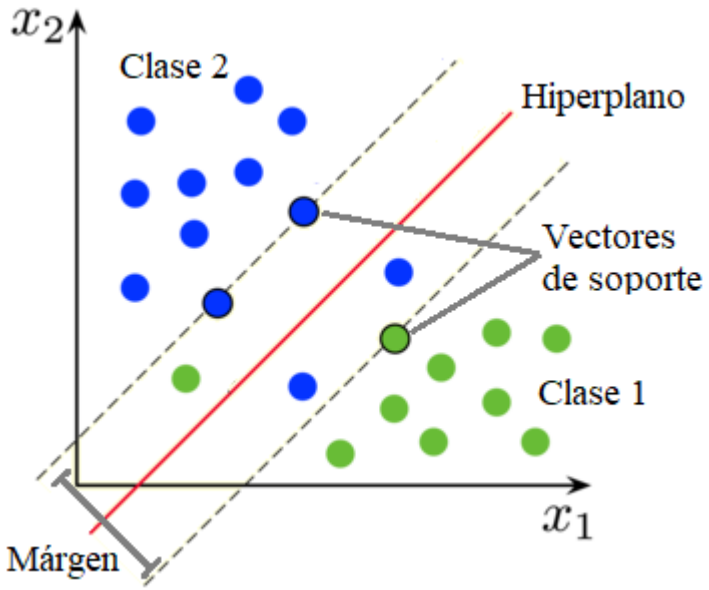

Figura 7. Máquina de Soporte Vectorial.

Para generar el umbral de decisión, la máquina de soporte vectorial cuenta con un margen definido por datos denominados vectores de soporte. Los vectores de soporte se encuentran al margen del hiperplano de separación y establecen una región donde se permite que el modelo realice clasificaciones posiblemente incorrectas, esto se denomina margen blando, en contraposición con un margen duro, que no admite ninguna clasificación incorrecta. La orientación del hiperplano de separación se optimiza de modo que maximice su distancia a los vectores de soporte y minimice la cantidad de errores en la clasificación.

3.8 Medidas de desempeño

A la hora de poner a prueba un modelo se usan estrategias para determinar la efectividad de este en la resolución del problema. Algunas medidas importantes son: Matriz de confusión, sensibilidad, especificidad, tasa de acierto, curva ROC, área bajo la curva (AUC).

3.8.1 Matriz de confusión.

La matriz de confusión identifica de forma ordenada la cantidad de datos que fueron clasificados correctamente en una clase y otra. Las dimensiones de la matriz de confusión están

determinadas por el cuadrado del número de clases del problema. Para comprender los conceptos asociados a una matriz de confusión de dos clases se identifica la primera como 'Positivo' y la segunda como 'Negativo'.

		Clase estimada	
		P	N
Clase Verdadera	P	TP	FN
	N	FP	TN

Tabla 1. Matriz de confusión biclase: Positivo (P) y Negativo (N).

En una matriz de confusión se clasifican los resultados de acuerdo con los resultados entregados por un modelo. Cada fila corresponde a las etiquetas reales de los datos de prueba, mientras que cada columna indica la etiqueta predicha por el modelo. Los datos en cada celda de la matriz son [27],[28]:

- **TN**, *True Negatives*: Datos de la clase negativa clasificados correctamente.
- **FP**, *False Positives*: Datos de la clase negativa clasificados incorrectamente.
- **FN**, *False Negatives*: Datos de la clase positiva clasificados incorrectamente.
- **TP**, *True Positives*: Datos de la clase positiva clasificados correctamente.

3.8.2 Tasa de acierto, Especificidad, Sensibilidad, F1 Score

A partir de la matriz de confusión, se pueden calcular algunas métricas importantes:

Tasa de acierto (CCR): cociente entre datos clasificados correctamente y número total de datos. Es una métrica general del desempeño del sistema, es apropiada para datos balanceados.

$$CCR = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

Sensibilidad (S): cociente entre verdaderos positivos y número total de positivos. Indica la capacidad del sistema de clasificar correctamente los datos de la clase positiva.

$$S = \frac{TP}{TP + FN} \quad (9)$$

Especificidad (E): cociente entre verdaderos negativos y número total de negativos. Indica la capacidad del sistema de clasificar correctamente los datos de la clase negativa.

$$E = \frac{TN}{TN + FP} \quad (10)$$

F1 Score: Para definir el F1 *Score* es necesario definir el parámetro *precisión (P)* mostrado en la *ecuación 11*. De esta manera, el F1 *Score* se define como la media armónica (promedio) entre la precisión y la sensibilidad, como se muestra en la *ecuación 12*. Indica el desempeño general del sistema.

$$P = \frac{TP}{TP + FP} \quad (11)$$

$$F1\ Score = 2 * \frac{S * P}{S + P} \quad (12)$$

3.8.3 Curva ROC

La curva ROC (del inglés *Receiver Operating Characteristic*) es una representación gráfica de la especificidad y la sensibilidad de un modelo. La curva ROC se obtiene cuando se grafica el número de verdaderos positivos en función de los falsos positivos cuando el umbral de decisión se mueve progresivamente de clasificar todos los datos como negativos (umbral a la izquierda) a clasificar todos los datos como positivos (umbral a la derecha) [29].

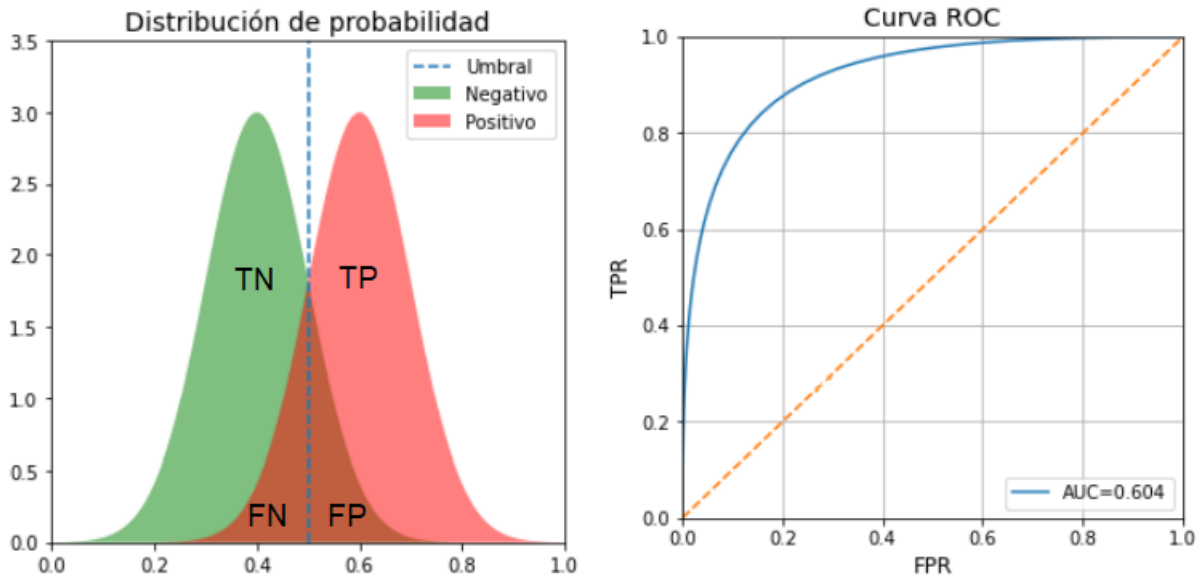

Figura 8. Construcción de la curva ROC.

3.8.4 Área bajo la curva ROC (AUC)

El área bajo la curva ROC brinda una cuantificación del desempeño general del sistema. El AUC tiene un valor máximo de 1 para un sistema con sensibilidad y especificidad ideales. Mientras que el valor mínimo de 0.5 indica que el sistema realiza clasificaciones aleatorias [29].

4 Base de Datos

Las bases de datos que se emplearán para el trabajo consisten en grabaciones de audio de hablantes nativos checos. Se tiene una base de datos por cada patología: Parkinson y Huntington. Una base de datos contiene 50 pacientes de Parkinson y 50 sujetos de control sanos [30] y otra contiene 40 pacientes de Huntington y 40 sujetos de control sanos [31]. Las grabaciones se realizaron en una habitación silenciosa con un micrófono a 5 cm de la boca del paciente [32]. Los pacientes de la base de datos de PD fueron evaluados por un neurólogo experto según la Escala Unificada para la Evaluación de la Enfermedad de Parkinson (UPDRS)[33] con una media de 19.8 en una escala entre 0 y 72, lo que indica que estos pacientes tenían una baja gravedad de las

manifestaciones motoras. Por otro lado, los pacientes de Huntington fueron evaluados según la Escala Unificada para la Evaluación de la Enfermedad de Huntington (UHDRS) [34]. El rango de la puntuación total de esta escala está entre 0 y 124, y la puntuación media de los pacientes de este corpus fue de 27.0, lo que indica que se encontraban en un estado temprano-intermedio de la enfermedad. Ninguno de los participantes sanos tenía antecedentes de trastornos neurológicos o de comunicación. La Tabla 1 resume la información de las bases de datos. De las pruebas estadísticas realizadas se encuentra que no hay evidencia para rechazar la hipótesis nula, es decir, que no hay diferencia entre la distribución estadística de pacientes y controles.

	Base de datos PD			**Base de datos HD**		
	PD	**HC**	**PD vs. HC**	**HD**	**HC**	**HD vs. HC**
Género [F/M]	20/30	20/30	$\boldsymbol{p} = 0.16$	20/20	20/20	$\boldsymbol{p} = 1.00$
Edad [F/M]	60±9/65±10	63±11/60±12	$\boldsymbol{*p} = 0.44$	49±14/48±12	50±14/48±12	$\boldsymbol{*p} = 0.43$
Rango de edad [F/M]	41-72/43-82	40-79/41-77		27-69/23-67	27-69/26-70	
UPDRS UHDRS [F/M]	18±10/21±12	– – –		27±10/27±13	– – –	

Tabla 2. Bases de datos en idioma checo. **PD**: Enfermedad de Parkinson, **HD**: Enfermedad de Huntington, **HC**: Individuos de control sanos. Los resultados se reportan como la media ± desviación estándar. **p:** valor p de prueba chi-cuadrado**, *p:** valor p de prueba Mann Whitney U.

5 Metodología

La metodología general planteada es resumida en la ***Figura 9***. Observando la gráfica de izquierda a derecha, se toman de las bases de datos de HD y PD segmentos de audio de 1 segundo de duración, los cuales servirán como las entradas del sistema. En primer lugar, las señales son analizadas mediante una CNN-1d compuesta por dos capas convolucionales y dos capas de *max pooling*. Cada capa convolucional contiene un conjunto de filtros entrenables en varios canales que buscan identificar características específicas de la señal en diferentes frecuencias. Particularmente se usa una CNN-1d para poder tratar las señales de audio en su representación original, la cual es unidimensional. Posteriormente, las salidas de la red neuronal convolucional son entregados a un conjunto de celdas LSTM. Considerando que una señal de audio es una serie de tiempo, el objetivo de la LSTM es interpretar la secuencialidad de las características encontradas por la CNN, modelando la interacción y las dependencias temporales entre los segmentos contiguos (o cercanos) de la señal original. Finalmente, las salidas de cada celda LSTM son concatenadas obteniendo un gran número de características, las cuales son reducidas gradualmente mediante varias capas neuronales completamente conectadas hasta obtener una clasificación binaria.

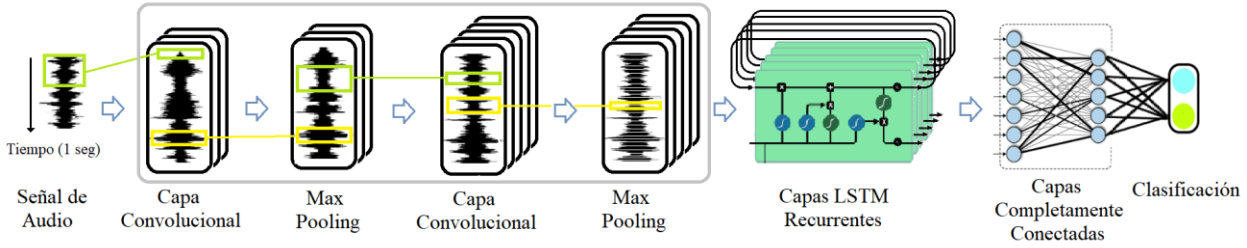

Figura 9. Metodología General

5.1 Experimentos

Con el objetivo de determinar los parámetros adecuados para la arquitectura planteada y evaluar su desempeño se plantean los siguientes experimentos: (i) *Baseline*, usando métodos clásicos para obtener resultados de referencia, (ii) *Diseño y optimización de la arquitectura*, donde se realiza la definición y variaciones del modelo, (iii) *Clasificación de disartrias*, realizando de forma independiente las clasificaciones para pacientes con disartria hipocinética e hipercinética

(asociada a la enfermedad de Parkinson y Huntington, respectivamente) vs. sus respectivos controles. Además, también se realiza la clasificación entre pacientes con disartria para la identificación del tipo de disartria exhibida por los pacientes con HD y PD.

5.1.1 Baseline

En primer lugar, como punto de referencia, se realizan una serie de experimentos de clasificación para las bases de datos de PD y HD usando métodos clásicos. Los escenarios de clasificación son los mismos que se realizarán para la arquitectura propuesta con redes neuronales, es decir: (i) Clasificación de pacientes con disartria hipocinética (PD) vs controles sanos, (ii) Clasificación de pacientes con disartria hipercinética (HD) vs controles sanos, (iii) Clasificación entre pacientes con disartria (Hipocinética vs Hipercinética). Específicamente se usan características de articulación y prosodia, tanto de forma independiente como combinada para representar la información de las señales de audio, luego una SVM es la encargada de clasificar cada uno de los escenarios propuestos. En todos los experimentos, se utiliza una estrategia de validación cruzada de 10 *folds* para garantizar que los resultados no dependen de una elección particular de conjuntos de entrenamiento y prueba.

5.1.2 Diseño y optimización de la arquitectura

El parámetro de interés que se busca optimizar en nuestra arquitectura es el tamaño de ventana analizado por los filtros de la CNN-1d. Para ello, se hace una partición del conjunto de datos con 70% de entrenamiento y 30% de prueba. Posteriormente el desempeño del sistema es evaluado modificando el tamaño del *kernel* (como se especifica en la sección 3.2) para analizar ventanas de 2.5 ms, 5 ms, 10 ms y 20 ms. Para este experimento no se aplica validación cruzada debido al elevado tiempo de computación requerido en tal proceso. Por otro lado, las variaciones en el desempeño del sistema para la misma partición del conjunto de datos son suficientes para determinar la influencia que tiene la elección particular del tamaño de ventana analizado por cada filtro de la CNN. Asimismo, se probaron distintos valores de *dropout* (0.1, 0.2, 0.3) para elegir el que diera lugar al mejor desempeño para proceder con la experimentación.

5.1.3 Clasificación de disartrias

Usando el tamaño de *kernel* encontrado en la anterior sección, se procede con los experimentos de clasificación de pacientes con disartria y controles sanos. La clasificación se realiza de forma independiente para la base de datos de pacientes con disartria hipocinética asociada a la enfermedad de Parkinson, como para la base de datos de pacientes con disartria hipercinética causada por la enfermedad de Huntington. Los experimentos son realizados con y sin *dropout* y usando una validación cruzada de 10 *folds*.

Además, se realiza la clasificación entre pacientes con disartria usando la misma estrategia del experimento anterior, con la diferencia de que la clasificación se realiza entre los pacientes con disartria hipocinética vs los pacientes con disartria hipercinética.

6 Resultados

6.1 Baseline

La ***tabla 3*** resume los resultados de los 3 escenarios de clasificación planteados utilizando métodos clásicos con las características de articulación, prosodia y la fusión temprana entre ellas. El mejor resultado para cada escenario de clasificación se encuentra sombreado en la tabla y en las siguientes subsecciones se presenta el análisis de cada escenario por separado. En general, se observa que la prosodia es un mejor indicador de la presencia de disartria en comparación con la articulación, lo cual indica que las enfermedades de PD y HD afectan mucho más el ritmo del habla de los pacientes, que la pronunciación particular de los fonemas que componen el habla.

Clasificación usando métodos clásicos (*Baseline*)					
Escenario	**Características**	**Tasa de Acierto ($\mu \pm \sigma$) %**	**Sensibilidad ($\mu \pm \sigma$) %**	**Especificidad ($\mu \pm \sigma$) %**	**F1 Score ($\mu \pm \sigma$) %**
PD vs HC 6.1.1	**Prosodia**	67 ± 16	66 ± 28	66 ± 13	65 ± 16
	Articulación	49 ± 10	48 ± 24	47 ± 13	46 ± 11
	Fusión	58 ± 18	59 ± 23	57 ± 27	56 ± 18

HD vs HC 6.1.2	**Prosodia**	86 ± 13	82 ± 20	90 ± 12	86 ± 13
	Articulación	67 ± 11	63 ± 18	73 ± 16	67 ± 11
	Fusión	87 ± 10	87 ± 15	87 ± 18	87 ± 10
HD vs PD 6.1.3	**Prosodia**	78 ± 16	78 ± 26	78 ± 21	76 ± 17
	Articulación	76 ± 13	74 ± 23	80 ± 14	75 ± 13
	Fusión	73 ± 10	77 ± 17	67 ± 22	71 ± 10

Tabla 3. Resultados de clasificación usando métodos clásicos (*Baseline*)

6.1.1 Clasificación de pacientes con disartria hipocinética (PD) vs controles sanos (HC)

La ***Figura 10*** muestra las curvas ROC obtenidas para la clasificación de pacientes con disartria hipocinética asociada a la enfermedad de Parkinson, usando características clásicas de prosodia, articulación y la fusión de las características. Las características de prosodia demuestran ser capaces de evidenciar la disartria de los pacientes hasta cierta medida, mostrando un área bajo la curva ROC de 0.72. Por otro lado, las características de articulación dan como resultado un desempeño similar a un clasificador binario aleatorio, es decir, para esta dimensión no se encuentran diferencias significativas entre el habla patológica (PD) y el habla sana (HC). A la derecha de la gráfica se muestra la distribución de los *scores* obtenidos con las características de prosodia, ya que estas dan como resultado a la mayor separación entre los datos. La gráfica de distribución evidencia que la separación entre las clases es baja, dando resultado a un gran número de errores en la clasificación y a tasas de acierto bajas. Por otro lado, tanto en la gráfica de distribución como en las curvas ROC se observa simetría respecto a la clasificación de ambas clases, indicando que el sistema no tiene sesgos hacia una clase en específico. Esto se evidencia también en el primer escenario de la ***Tabla 3,*** donde los valores de sensibilidad y especificidad son similares para cada experimento.

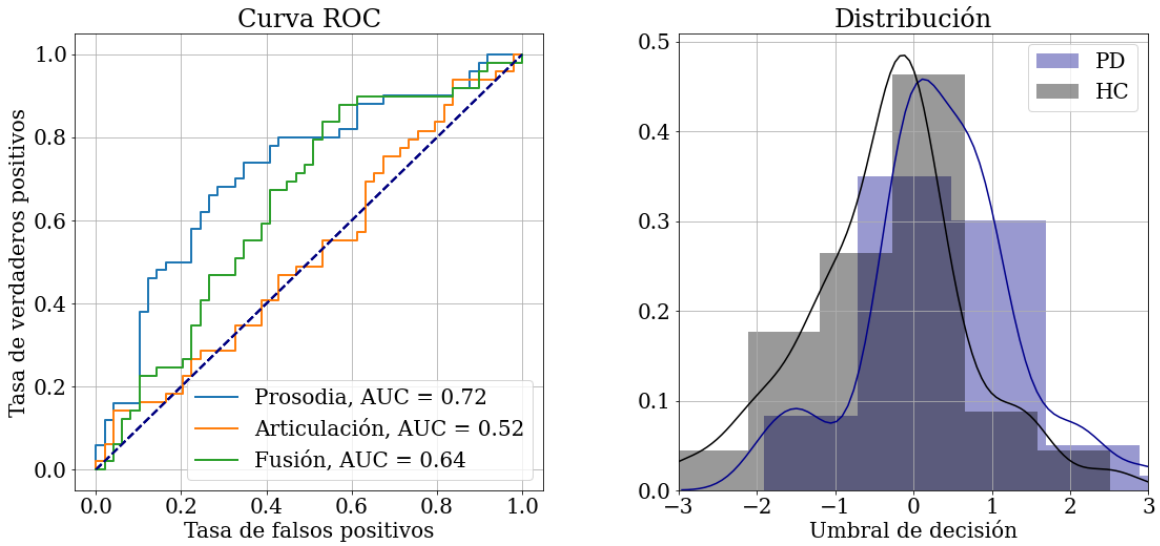

Figura 10. Curvas ROC y distribución para el mejor resultado, experimento de PD vs HC (*Baseline*)

6.1.2 Clasificación de pacientes con disartria hipercinética (HD) vs controles sanos (HC)

La ***Figura 11*** muestra las curvas ROC obtenidas para la clasificación de pacientes con disartria hipercinética asociada a la enfermedad de Huntington, usando características clásicas de prosodia, articulación y la fusión de las características. Para este escenario, los mejores resultados se encuentran usando la fusión de características de prosodia y articulación con un AUC de 0.96. Como se ha mencionado anteriormente, la mayor parte de la información para distinguir el habla patológica está dada por las características de prosodia con un AUC de 0.93. Sin embargo, en este caso las características de articulación también son capaces de modelar y aportar información no capturada con las características de prosodia. Se puede evidenciar que los síntomas de los pacientes de HD son mucho más pronunciados que para PD, posiblemente por un mayor avance de la enfermedad. Además, en la gráfica de la distribución se observa que existe una separación considerable entre las clases, dando lugar a valores altos para la tasa de acierto en la clasificación. No se observa ningún sesgo marcado en la distribución y la ***Tabla 3*** corrobora que el valor de sensibilidad y especificidad son idénticos.

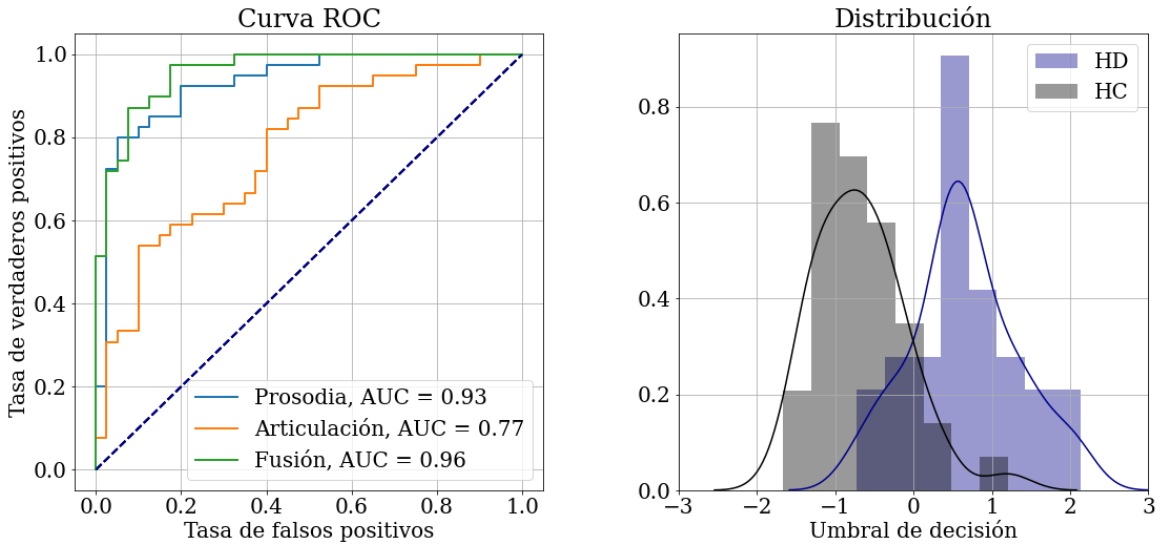

Figura 11. Curvas ROC y distribución para el mejor resultado, experimento de HD vs HC (*Baseline*)

6.1.3 Clasificación entre pacientes con disartria (Disartria Hipocinética vs Hipercinética)

Por último, la ***Figura 12*** muestra las curvas ROC obtenidas para la clasificación de pacientes con disartria hipercinética asociada a la enfermedad de Huntington vs. pacientes con disartria hipocinética causada por la enfermedad de Parkinson y la gráfica de distribución del mejor resultado, correspondiente al hallado usando características de prosodia. A diferencia de los dos casos anteriores, las características de articulación muestran ser capaces de modelar las variaciones entre la disartria hipocinética e hipercinética obteniendo un rendimiento similar a las características de prosodia. Las 3 curvas ROC presentan valores AUC similares, sobresaliendo el área bajo la curva obtenida a partir de las características de Prosodia. Además, a pesar de presentarse un solapamiento grande en la distribución de ambas clases, la separación entre las mismas es suficiente para obtener una tasa de acierto de 78% como se observa en la sección final de la ***Tabla 3***.

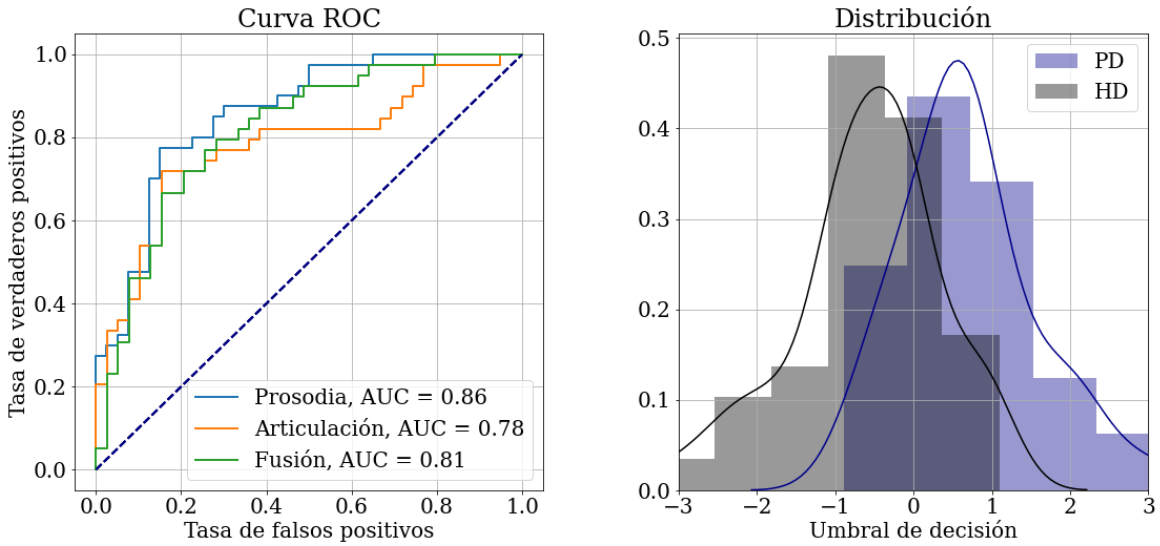

Figura 12. Curvas ROC y distribución para el mejor resultado, experimento de HD vs PD (*Baseline*)

6.2 Diseño y optimización de la arquitectura

La ***Figura 13*** muestra el desempeño de los modelos de clasificación para las bases de datos de Huntington y Parkinson respectivamente, usando tamaños de ventana de entre 2.5 ms hasta 40 ms. En ambos casos, se encuentra que la longitud de ventana más apropiado para los filtros de la CNN-1d es de 20ms. Este es el valor, es usado para todos los modelos siguientes.

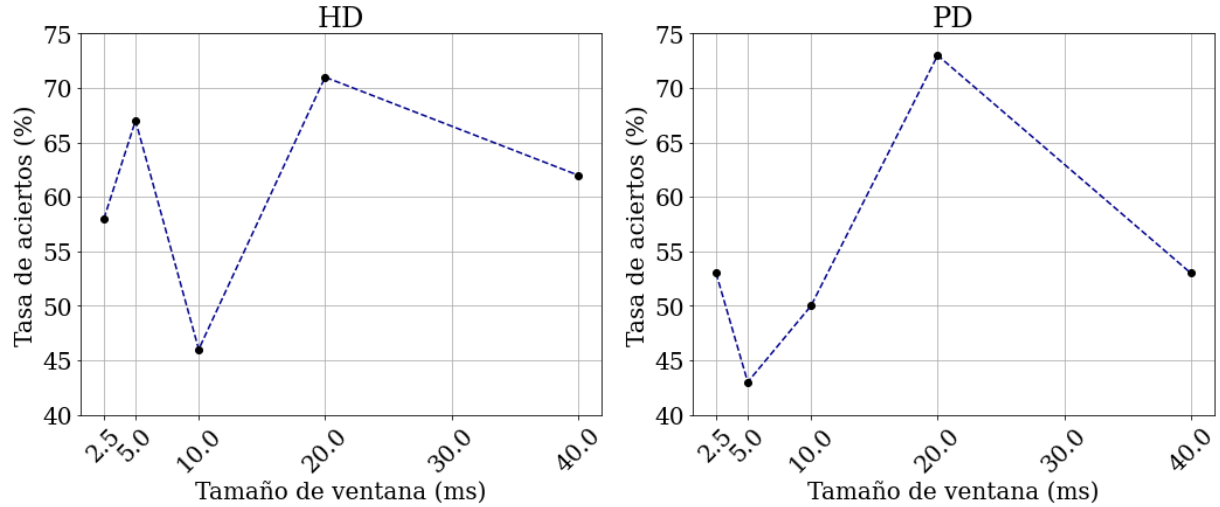

Figura 13. Experimentos con variación del tamaño de ventana en los filtros de CNN-1d

En la ***Tabla 4*** se muestra el desempeño de los modelos de clasificación usando diferentes niveles de *dropout*: 0.1, 0.2 y 0.3, y un tamaño de ventana de 20 ms. Para la base de datos de HD, el mejor resultado se encuentra con un *dropout* de 0.3 mientras que, para la base de datos de PD, el *dropout* es de 0.1. Estos valores de *dropout* son los usados en los experimentos de clasificación con una estrategia de validación cruzada.

Experimentación de *Dropout*					
Base de datos	***Dropout***	**Tasa de Acierto**	**Sensibilidad**	**Especificidad**	**F1 Score**
PD	**0.1**	70 %	64 %	75 %	70 %
	0.2	60 %	86 %	38 %	59 %
	0.3	53 %	86 %	25 %	50 %
HC	**0.1**	71 %	91 %	54 %	71 %
	0.2	67 %	45 %	85 %	65 %
	0.3	71 %	64 %	77 %	71 %

Tabla 4. Experimentación de *Dropout*

6.3 Clasificación de disartrias y controles sanos

La ***Tabla 5*** recopila los resultados de los 3 escenarios de clasificación planteados utilizando la arquitectura de aprendizaje profundo planteada y una estrategia de validación cruzada con 10 *folds*. El mejor resultado para cada escenario de clasificación se encuentra sombreado en la tabla y en las siguientes subsecciones se presenta el análisis de cada escenario. La desviación estándar de los resultados es alta debido a la reducida cantidad de datos e iteraciones posibles considerando el alto tiempo de computación necesario para cada experimento. Duranta cada iteración de la validación cruzada, el conjunto de prueba contiene 10 o menos elementos, resultando en un error de 10% por cada clasificación errónea.

Clasificación usando Aprendizaje Profundo					
Escenario	***Dropout***	**Tasa de Acierto ($\mu \pm \sigma$) %**	**Sensibilidad ($\mu \pm \sigma$) %**	**Especificidad ($\mu \pm \sigma$) %**	**F1 Score ($\mu \pm \sigma$) %**

PD vs HC 6.3.1	**0.0**	65 ± 10	84 ± 15	47 ± 15	64 ± 11
	0.1	63 ± 18	74 ± 27	52 ± 22	61 ± 18
HD vs HC 6.3.2	**0.0**	71 ± 13	78 ± 26	65 ± 23	70 ± 13
	0.3	69 ± 19	70 ± 29	68 ± 32	66 ± 23
HD vs PD 6.3.3	**0.0**	71 ± 20	70 ± 38	72 ± 26	67 ± 25
	0.1	74 ± 10	72 ± 28	75 ± 19	71 ± 14

Tabla 5. Resultados de clasificación usando aprendizaje profundo

6.3.1 Clasificación de pacientes con disartria hipocinética (PD) vs controles sanos (HC)

La ***Figura 14*** muestra las curvas ROC para los experimentos de clasificación de disartria hipocinética (PD) vs controles sanos (HC) con y sin *dropout*. La gráfica de distribución muestra una separación baja entre las dos clases, dando lugar a un limitado desempeño en la clasificación. Además, la gráfica muestra un sesgo hacia la clase positiva (PD), resultando en un valor de sensibilidad mayor al valor de especificidad, como se evidencia en la ***Tabla 5***. Además, en la tabla se observa que la tasa de acierto obtenida (63% ± 18%) se encuentra muy cercana al valor obtenido en el *baseline*.

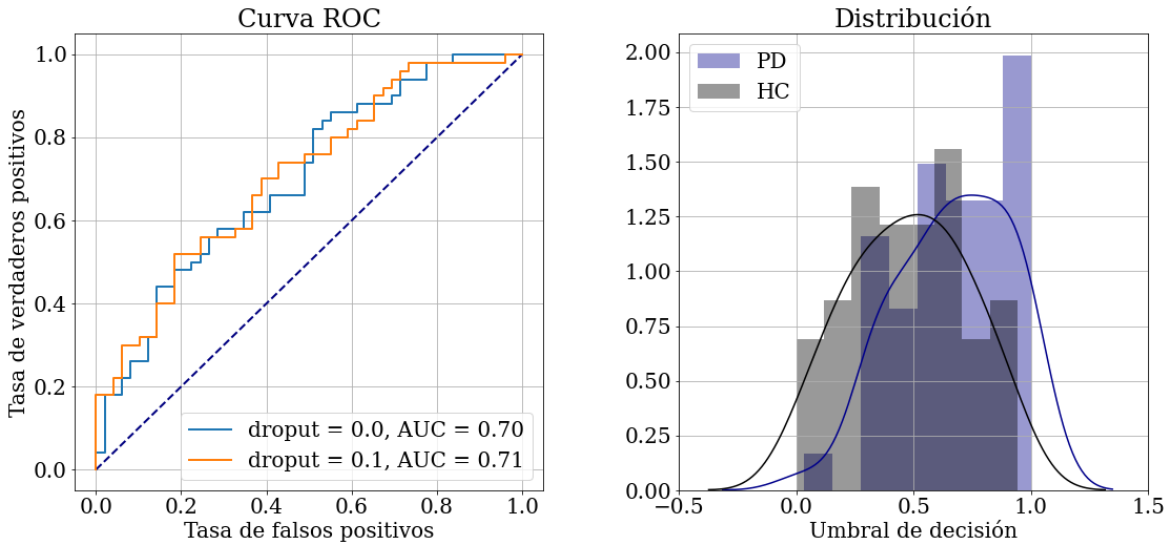

Figura 14. Curvas ROC y distribución para el mejor resultado, experimento de PD vs HC

6.3.2 Clasificación de pacientes con disartria hipercinética (HD) vs controles sanos (HC)

Al igual que en el *baseline*, la clasificación de disartria hipercinética asociada a HD con el habla sana (HC) presenta mejores resultados que la clasificación de disartria hipocinética asociada a PD. Sin embargo, el desempeño obtenido con la arquitectura propuesta es menor que el *baseline*. Al igual que la sección anterior, la gráfica de la distribución se observa que el clasificador identifica más fácilmente los pacientes que los controles, dando como resultado una sensibilidad mayor que la especificidad. En este caso, el área bajo las curvas ROC indica que el experimento sin *dropout* brinda resultados ligeramente mejores.

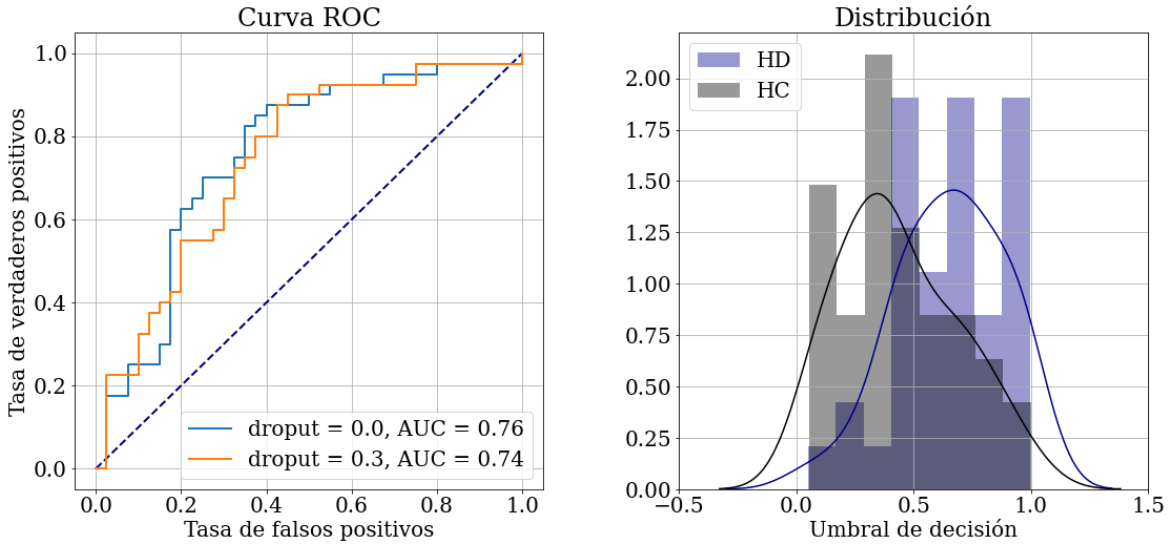

Figura 15. Curvas ROC y distribución para el mejor resultado, experimento de HD vs HC

6.3.3 Clasificación entre pacientes con disartria (Disartria Hipocinética vs Hipercinética)

Finalmente, la ***Figura 16*** muestra las curvas ROC y la distribución de *scores* para el experimento de clasificación entre pacientes con disartria Hipercinética asociada a HD y disartria Hipocinética asociada a PD. En la gráfica de distribución se observa que se identifican más fácilmente los datos de la clase negativa (HC) y la ***Tabla 5*** confirma que la especificidad es mayor que la sensibilidad. La separación entre las clases es considerable y da como resultado una tasa de

acierto de 74% ± 10% cuando se usa un valor de *dropout* de 0.1. Los resultados obtenidos para este experimento son similares a los obtenidos con métodos clásicos en el *baseline*.

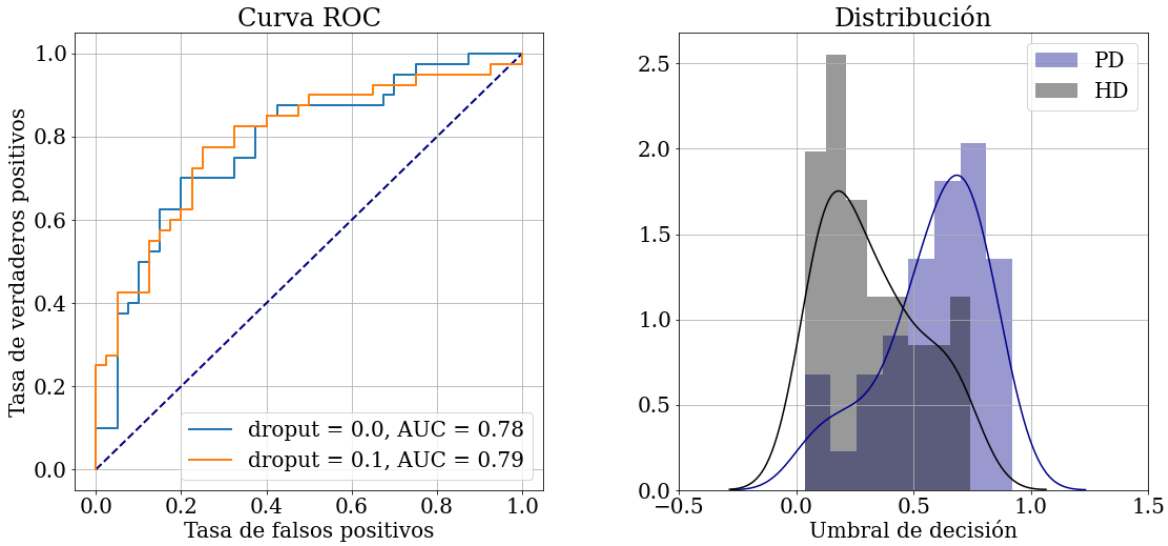

Figura 16. Curvas ROC y distribución para el mejor resultado, experimento de HD vs PD

7 Conclusiones

En el presente trabajo se diseñó una arquitectura de redes neuronales convolucionales seguidas de celdas LSTM para la clasificación de señales de voz de pacientes de enfermedades neurodegenerativas, específicamente HD y PD (las cuales generalmente están asociadas a condiciones denominadas disartria hipercinética y disartria hipocinética, respectivamente) y señales de voz de individuos sanos (HC). Se utilizó una metodología sistemática para determinar los parámetros de la arquitectura y evaluar el desempeño del modelo para la clasificación binaria en diferentes escenarios: PD vs HC, HD vs HC y HD vs PD. Además, se implementó un modelo de referencia (*baseline*) usando métodos clásicos de aprendizaje de máquina y características extraídas manualmente con el objetivo de comparar ambas metodologías.

Para el modelo de referencia se obtuvo un 67% ± 16% de tasa de acierto en la clasificación de disartria hipocinética y voz sana (PD vs HC), 87% ± 10% en la clasificación de disartria hipercinética y voz sana (HD vs HC), y 78% ± 16% en la clasificación de disartria hipocinética y disartria hipercinética (PD vs HD). Se evidenció que el mejor indicador para ambos tipos de disartria es la prosodia, relacionada principalmente con el ritmo natural de habla. Esto concuerda con la descripción de la disartria hipocinética, asociada a monotonía en la voz y la disartria hipercinética caracterizada por una cadencia anormal y aleatoria del habla. Por otro lado, con el modelo de aprendizaje profundo se reportan resultados de 63% ± 18% en la tasa de acierto en la clasificación de disartria hipocinética y voz sana (PD vs HC), 71% ± 13% en la clasificación de disartria hipercinética y voz sana (HD vs HC), y 74% ± 10% en la clasificación de disartria hipocinética y disartria hipercinética (PD vs HD). En general los resultados del modelo de aprendizaje profundo se encuentran ligeramente por debajo de los resultados obtenidos con el modelo de referencia, exceptuando el caso de la clasificación entre (HD vs HC), donde el modelo de referencia supera ampliamente al modelo de aprendizaje profundo. Aunque esto se puede explicar en cierta medida por la efectividad inherente de usar la prosodia como una característica en el modelo de referencia y una fase avanzada de la enfermedad de Huntington, también es cierto que el modelo de aprendizaje profundo puede requerir de más datos, experimentación y optimización de parámetros para mejorar su desempeño. Sin embargo, el tiempo de computación necesario para la realización de una experimentación más grande y exhaustiva excede el alcance de este trabajo.

A pesar de que los resultados del modelo de aprendizaje profundo no mostraron ninguna ventaja sobre el modelo clásico de referencia, el marco de trabajo sistemático para el diseño y experimentación de modelos de aprendizaje profundo desarrollado es útil y se puede utilizar en el futuro para evaluar y optimizar modelos cuando no se cuente características clásicas que den lugar a un buen desempeño en la clasificación (como es el caso de la prosodia como marcador de disartria). Además, se evidencia que los métodos clásicos de aprendizaje de máquina no pueden ser pasados por alto y de estar disponibles, son frecuentemente un buen punto de partida mientras que los modelos de aprendizaje profundo pueden llegar a presentar mejor resultados, aunque con un costo computacional mucho mayor.

Referencias

REFERENCIAS BIBLIOGRÁFICAS

[1] Schapira, A. H., Olanow, C. W., Greenamyre, J. T., & Bezard, E. (2014). Slowing of neurodegeneration in Parkinson's disease and Huntington's disease: future therapeutic perspectives. The Lancet, 384(9942), 545-555.

[2] Jankovic, J. (2008). Parkinson's disease: clinical features and diagnosis. Journal of neurology, neurosurgery & psychiatry, 79(4), 368-376.

[3] Walker, F. O. (2007). Huntington's disease. The Lancet, 369(9557), 218-228.

[4] Pinto, S., Ozsancak, C., Tripoliti, E., Thobois, S., Limousin-Dowsey, P., & Auzou, P. (2004). Treatments for dysarthria in Parkinson's disease. The Lancet Neurology, 3(9), 547-556.

[5] Hartelius, L., Carlstedt, A., Ytterberg, M., Lillvik, M., & Laakso, K. (2003). Speech disorders in mild and moderate Huntington disease: Results of dysarthria assessments of 19 individuals. Journal of Medical Speech-Language Pathology, 11(1), 1-15.

[6] Lowit, A., Marchetti, A., Corson, S., & Kuschmann, A. (2018). Rhythmic performance in hypokinetic dysarthria: Relationship between reading, spontaneous speech and diadochokinetic tasks. Journal of Communication Disorders, 72, 26.

[7] Rusz, J., Tykalova, T., Ramig, L. O., & Tripoliti, E. (2021). Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. Movement Disorders, 36(4), 803-814.

[8] Novotný, M., Rusz, J., Čmejla, R., Růžičková, H., Klempíř, J., & Růžička, E. (2016). Hypernasality associated with basal ganglia dysfunction: evidence from Parkinson's disease and Huntington's disease. PeerJ, 4, e2530.

[9] Hlavnička, J., Tykalová, T., Ulmanová, O., Dušek, P., Horáková, D., Růžička, E., ... & Rusz, J. (2020). Characterizing vocal tremor in progressive neurological diseases via automated acoustic analyses. Clinical Neurophysiology, 131(5), 1155-1165.

[10] Moro-Velazquez, L., Gomez-Garcia, J. A., Godino-Llorente, J. I., Villalba, J., Rusz, J., Shattuck-Hufnagel, S., & Dehak, N. (2019). A forced gaussians based methodology for the differential evaluation of Parkinson's Disease by means of speech processing. Biomedical Signal Processing and Control, 48, 205-220.

[11] Solana-Lavalle, G., Galán-Hernández, J. C., & Rosas-Romero, R. (2020). Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features. Biocybernetics and Biomedical Engineering, 40(1), 505-516.

[12] Vásquez-Correa, J. C., Rios-Urrego, C. D., Arias-Vergara, T., Schuster, M., Rusz, J., Nöth, E., & Orozco-Arroyave, J. R. (2021). Transfer learning helps to improve the accuracy to classify patients with different speech disorders in different languages. Pattern Recognition Letters, 150, 272-279.

[13] Vásquez-Correa, J. C., Arias-Vergara, T., Rios-Urrego, C. D., Schuster, M., Rusz, J., Orozco-Arroyave, J. R., & Nöth, E. (2019). Convolutional neural networks and a transfer learning strategy to classify Parkinson's disease from speech in three different languages. In Iberoamerican Congress on Pattern Recognition (pp. 697-706). Springer, Cham.

[14] Mian, T. S. (2022). An Unsupervised Neural Network Feature Selection and 1D Convolution Neural Network Classification for Screening of Parkinsonism. Diagnostics, 12(8), 1796.

[15] Rios-Urrego, C.D., Moreno-Acevedo, S.A., Nöth, E., Orozco-Arroyave, J.R. (2022). End-to-End Parkinson's Disease Detection Using a Deep Convolutional Recurrent Network. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds) Text, Speech, and Dialogue. TSD 2022. Lecture Notes in Computer Science(), vol 13502. Springer, Cham.

[16] Mallela, J., Illa, A., Suhas, B. N., Udupa, S., Belur, Y., Atchayaram, N., ... & Ghosh, P. K. (2020). Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson's Disease and healthy controls with CNN-LSTM using transfer learning. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6784-6788

[17] Orozco-Arroyave, J. R., Vásquez-Correa, J. C., Vargas-Bonilla, J. F., Arora, R., Dehak, N., Nidadavolu, P. S., ... & Nöth, E. (2018). NeuroSpeech: An open-source software for Parkinson's speech analysis. Digital Signal Processing, 77, 207-221.

[18] Dehak, N., Dumouchel, P., & Kenny, P. (2007). Modeling prosodic features with joint factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing, 15(7), 2095-2103.

[19] S. Saha. (2018) A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way. Toronto, Ontario: Towards Data Science. Recuperado de: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

[20] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 8024–8035

[21] Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. Mechanical systems and signal processing, 151, 107398.

[22] Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), 157-166.

[23] Shenfield, A., & Howarth, M. (2020). A novel deep learning model for the detection and identification of rolling element-bearing faults. Sensors, 20(18), 5112.

[24] Er, M. B., Isik, E., & Isik, I. (2021). Parkinson's detection based on combined CNN and LSTM using enhanced speech signals with variational mode decomposition. Biomedical Signal Processing and Control, 70, 103006.

[25] A. Nagpal (2017) L1 and L2 Regularization Methods. Towards Data Science. Recuperado de: https://towardsdatascience.com/l1-and-l2-regularization-methods-ce25e7fc831c

[26] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[27] A. Mitrani. (2019) Evaluating Categorical Models II: Sensitivity and Specificity. Toronto, Ontario: Towards Data Science. Recuperado de: https://towardsdatascience.com/evaluating-categorical-models-ii-sensitivity-and-specificity-e181e573cff8

[28] A. Mitrani. (2019) Evaluating Categorical Models. Toronto, Ontario: Towards Data Science. Recuperado de: https://towardsdatascience.com/evaluating-categorical-models-e667e17987fd

[29] D. Steen. (2020) Understanding the ROC Curve and AUC. Toronto, Ontario: Towards Data Science. Recuperado de: https://towardsdatascience.com/understanding-the-roc-curve-and-auc-dd4f9a192ecb

[30] J. Rusz. (2018) Detecting speech disorders in early Parkinson's disease by acoustic analysis. Habilitation thesis, Czech Technical University in Prague.

[31] Rusz, J., Klempíř, J., Tykalová, T., Baborová, E., Čmejla, R., Růžička, E., & Roth, J. (2014). Characteristics and occurrence of speech impairment in Huntington's disease: possible influence of antipsychotic medication. Journal of Neural Transmission, 121(12), 1529-1539.

[32] Rusz, J., Tykalova, T., Ramig, L. O., & Tripoliti, E. (2021). Guidelines for speech recording and acoustic analyses in dysarthrias of movement disorders. Movement Disorders, 36(4), 803-814.

[33] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. (2003). The unified Parkinson's disease rating scale (UPDRS): status and recommendations. Movement Disorders, 18(7), 738-750.

[34] Kremer, H. P. H., & Hungtington Study Group. (1996). Unified Huntington's disease rating scale: reliability and consistency. Movement disorders, 11, 136-142.