CARACTERIZACIÓN DEL TEMBLOR DE MANOS EN PACIENTES CON PARKINSON UTILIZANDO UN ESQUEMA DE APRENDIZAJE CONVOLUCIONAL PROFUNDO

JESSICA FERNANDA PEDRAZA CADENA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA

2022

CARACTERIZACIÓN DEL TEMBLOR DE MANOS EN PACIENTES CON PARKINSON UTILIZANDO UN ESQUEMA DE APRENDIZAJE CONVOLUCIONAL PROFUNDO

JESSICA FERNANDA PEDRAZA CADENA

Una tesis presentada en cumplimiento de los requisitos para el grado de:
Ingeniera de Sistemas e Informática

Director:
Fabio Martínez Carrillo
Ph.D en Ingeniería de Sistemas y Computación

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FÍSICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2022

AGRADECIMIENTOS

El autor expresa su agradecimiento:

Principalmente a Dios y a mis padres, Roman Pedraza y Gladys Cadena, quienes han sido un apoyo muy importante a lo largo de mi vida y un pilar fundamental para el cumplimiento de mis metas y propósitos.

A mi hermana Neila quien ha sido mi ejemplo a seguir, quien también ha sido un soporte importante en mi vida, quien por medio de sus consejos me ayuda a tomar buenas decisiones, quien ha estado para mi en mis momentos de dificultad y siempre que lo necesito.

A mi director de proyecto, Fabio Martinez, por su infinita paciencia a lo largo de los años, por creer en mi y enseñarme que los grandes esfuerzos conllevan grandes resultados, por ser ese gran docente y guía en todo este proceso.

Al ingeniero John Archila que estuvo desde el inicio en este proyecto, apoyándome y brindándome todo su conocimiento, por ser quién me impulso a seguir adelante, a no dejarme vencer, por sus consejos y por su gran paciencia.

A mis amigos tanto de la vida como de la universidad, quienes de alguna u otra forma han estado para mi y han aportado en todo este proceso dándome consejos y buenos momentos, así como apoyo incondicional.

Finalmente, a mis compañeros del grupo de investigación $BIVL^2ab$, principalmente a Alejandra y a Franklin, quienes por medio de su conocimiento y de sus consejos, han sido escalones fundamentales en esta investigación.

Índice general

Pág

Índice de figuras

Pág

Índice de cuadros

Pág

RESUMEN

TÍTULO: CARACTERIZACIÓN DEL TEMBLOR DE MANOS EN PACIENTES CON PARKINSON UTILIZANDO UN ESQUEMA DE APRENDIZAJE CONVOLUCIONAL PROFUNDO [*]

AUTOR: JESSICA FERNANDA PEDRAZA CADENA [**]

PALABRAS CLAVE: Enfermedad de Parkinson, temblor postural, temblor en reposo, análisis de video, representaciones profundas, redes convolucionales.

DESCRIPCIÓN: Actualmente, más de 6 millones de personas alrededor del mundo padecen la enfermedad de Parkinson (EP) y se estima que para el año 2040 el número de personas diagnosticadas ascienda a 17 millones. Este trastorno neurodegenerativo está relacionado con el déficit de dopamina, afectando principalmente condiciones motoras, tales como: lentitud de los movimientos, inestabilidad postural, temblor en las extremidades, rigidez, disminución en la amplitud del movimiento, afectaciones en la expresión facial y en la voz. El temblor siendo un movimiento rítmico y no controlado es el síntoma de mayor prevalencia en la EP afectando principalmente las extremidades. En la rutina clínica, la valoración y cuantificación de la enfermedad se puede lograr mediante la detección y caracterización del temblor en las manos siguiendo esquemas posturales y de reposo. En la configuración de reposo, las manos descansan sobre una superficie, limitando la percepción del temblor, especialmente en estadios tempranos. Por otra parte, en la configuración postural el paciente mantiene perpendicularmente sus brazos respecto al tronco, de esta manera, la amplitud del temblor se amplifica debido a la fuerza gravitacional. Este tipo de configuración, sin embargo, añade contracciones musculares voluntarias, resultando en señales ruidosas respecto a la caracterización motora propia del temblor. Estas valoraciones son además subjetivas y dependen de la experticia de los profesionales para determinar si el temblor está asociado a la EP. Hoy en día, en la literatura se reportan alternativas para la cuantificación del temblor, siendo un soporte para la caracterización de este patrón. Sin embargo, estas herramientas por lo general son invasivas y reportan una descripción limitada de los patrones motores del temblor.

En este trabajo se presenta una representación profunda volumétrica para la caracterización de los patrones de temblor asociados a la EP, registrados en secuencias de video bajo esquemas de reposo y postural. La estrategia

---

[*] Trabajo de investigación

[**] Facultad de Ingenierías Físico-Mecánicas. Escuela de Ingeniería de Sistemas e Informática. Director: Fabio Martínez Carrillo, Ph.D.

incluye un esquema convolucional que extrae patrones espacio-temporales, correlacionados con el temblor, los cuales son propagados a través de una arquitectura jerárquica que se ajusta respecto a una regla de discriminación entre pacientes con la EP y sujetos control. Además, se logró extraer características aprendidas en la representación profunda, permitiendo generar mapas de atención que permiten establecer las principales regiones temporales que se destacan en el video para realizar la clasificación y servir como soporte en la caracterización de la enfermedad. El método fue evaluado sobre un conjunto total de 80 videos (5 pacientes con la EP y 5 sujetos control), utilizando un esquema de validacion cruzada de tipo "leave one patient out", es decir, dejando un paciente fuera para evaluación y el resto para entrenamiento. En está validación se reporto una exactitud promedio de 92.5 % y una sensibilidad de 100 % en un esquema de reposo. En cuanto a un esquema postural, el método propuesto logra una exactitud promedio de 90 % y una sensibilidad de 80 %.

ABSTRACT

TITLE: CHARACTERIZATION OF HAND TREMOR IN PARKINSON'S PATIENTS USING A DEEP CONVOLUTIONAL LEARNING SCHEME [*]

AUTHOR: JESSICA FERNANDA PEDRAZA CADENA [**]

KEYWORDS: Parkinson's disease, tremor, postural tremor, resting tremor, video sequences, deep representations, convolutional networks.

DESCRIPTION: Currently, more than 6 million people around the world suffer from Parkinson's disease (PD) and it is estimated that by the year 2040 the number of diagnosed people will reach 17 million. This neurodegenerative disorder is related to dopamine deficiency, affecting mainly motor conditions, such as: slowness of movement, postural instability, tremor in the limbs, rigidity, decreased range of motion, and impairment of facial expression and voice. Tremor, being a rhythmic and uncontrolled movement, is the most prevalent symptom in PD, affecting mainly the extremities. In routine clinical practice, assessment and quantification of the disease can be achieved by detecting and characterizing tremor in the hands following postural and resting patterns. In the resting configuration, the hands rest on a surface, limiting tremor perception, especially in early stages. On the other hand, in the postural configuration the patient keeps his arms perpendicular to the trunk, thus, the amplitude of the tremor is amplified due to the gravitational force. This type of configuration, however, adds voluntary muscle contractions, resulting in noisy signals with respect to the proper motor characterization of the tremor. These assessments are also subjective and depend on the expertise of professionals to determine whether the tremor is associated with PD. Nowadays, alternatives for the quantification of tremor are reported in the literature, being a support for the characterization of this pattern. However, these tools are generally invasive and report a limited description of tremor motor patterns.

In this work we present a volumetric deep representation for the characterization of tremor patterns associated with PD, recorded in video sequences under resting and postural schemes. The strategy includes a convolutional scheme that extracts spatio-temporal patterns, correlated with tremor, which are propagated through a hierarchical architecture that is adjusted with respect to a discrimination rule between PD patients and control

---

[*] Research work

[**] Faculty of Physical-Mechanical Engineering. School of Systems and Computer Engineering. Advisor: Fabio Martínez Carrillo

subjects. In addition, it was possible to extract features learned in the deep representation, allowing to generate attention maps that allow to establish the main temporal regions that stand out in the video to perform the classification and serve as a support in the characterization of the disease. The method was evaluated on a total set of 80 videos (5 PD patients and 5 control subjects), using a cross-validation scheme of the "leave a patient out"type, that is, leaving one patient out for evaluation and the rest for training. In this validation, an average accuracy of 92.5 % and a sensitivity of 100 % in a resting scheme. As for a postural scheme, the proposed method achieves an average accuracy of 90 % and a sensitivity of 80 %.

INTRODUCCIÓN

La enfermedad de Parkinson (EP) es la segunda enfermedad neurodegenerativa más prevalente en el mundo después del Alzheimer siendo actualmente el trastorno neurodegenerativo de más rápido crecimiento afectando a millones de personas en todo el mundo[1]. La EP afecta al 1 % de las personas mayores de 65 años y aproximadamente al 4 % de la población mayor de 80 años[2]. En Colombia hay una prevalencia estimada de 4.7 pacientes por cada 1000 habitantes detectándose con mayor frecuencia en personas mayores de 60 años[3]. Fisiológicamente, la EP se asocia con una pérdida progresiva de dopamina, un neurotransmisor encargado de desarrollar procesos de locomoción óptimos y sincronizados. De esta manera, este déficit produce las alteraciones motoras relacionadas a movimientos coordinados y de equilibrio.

La inestabilidad postural, rigidez muscular, lentitud en los movimientos, el temblor y la afectación de la voz, constituyen el conjunto de síntomas típicos de la EP. El temblor, como uno de los síntomas con mayor prevalencia, se define como un movimiento rítmico e involuntario causado por inervaciones recíprocas de un músculo que conduce a contracciones repetitivas[4]. La detección y caracterización de este síntoma establece mecanismos complementarios a los protocolos estandarizados permitiendo el diagnóstico de la enfermedad.

---

[1] Karin Wirdefeldt y col. "Epidemiology and etiology of Parkinson's disease: a review of the evidence". En: *European journal of epidemiology* 26.1 (2011), pág. 1.

[2] Sushil Sharma y col. "Biomarkers in Parkinson's disease (recent update)". En: *Neurochemistry international* 63.3 (2013), págs. 201-229.

[3] Aracelly Castro Toro y Omar Freddy Buriticá. "Parkinson's disease: diagnostic criteria, risk factors and progression, and assessment scales clinical stage". En: *Acta Neurológica Colombiana* 30.4 (2014), págs. 300-306.

[4] Rodger J Elble. "Tremor". En: *Neuro-geriatrics*. Springer, 2017, págs. 311-326.

Durante la rutina clínica, los patrones anormales del temblor asociados a la EP se evalúan siguiendo esquemas de reposo y postural. En la configuración de reposo, los brazos se apoyan sobre los músculos o sobre una superficie firme y se observa el temblor en relajación. Alternativamente, en la configuración postural, los brazos se mantienen extendidos en contra de la gravedad a un ángulo de 90° respecto al cuerpo. De esta manera, el esfuerzo físico que ocasiona mantener esta postura genera una exageración del movimiento. Esta sobrecarga, sin embargo, puede interferir con el análisis de temblores naturales y por ende, generar señales ruidosas que limiten la adecuada cuantificación de dicho patrón. Esta cuantificación usualmente se da mediante herramientas tecnológicas que implican el uso de dispositivos electrónicos y sistemas portátiles basados en sensores de tipo inercial, como acelerómetros y electromiogramas[5]. Estas herramientas son en muchos casos invasivas y requieren intervención manual, lo que altera la naturalidad de los movimientos en los pacientes dificultando la caracterización de la enfermedad. Por otra parte, en la configuración postural, estas heramientas difieren en las direcciones de aceleración y las variaciones de velocidad angulares asociadas al temblor, debido a la implicacion de los sensores y alteraciones propias de los mecanismos de captura.

Este trabajo introduce una representación volumétrica convolucional con la capacidad de codificar patrones espacio-temporales relacionados al temblor de las manos permitiendo discriminar entre una población parkinsoniana y una población control. La estrategia propuesta discrimina entre estas poblaciones usando configuraciones en reposo y configuraciones posturales, brindando una alternativa para el soporte clínico en cuanto a la cuantificación y caracterización de la enfermedad. Para ello, se utilizaron secuencias de video magnificadas, es decir, secuencias que usan una descomposición óptica que permite resaltar las frecuencias de movimiento con mayor relación al fenómeno del temblor. Estas secuencias son mapeadas a una arquitectura volumétrica convolucional cuya principal característica es la recuperación de patrones temporales con

---

[5] Paulo Henrique G Mansur y col. "A review on techniques for tremor recording and quantification". En: *Critical Reviews™ in Biomedical Engineering* 35.5 (2007).

diferentes escalas temporales y espaciales.

Esta codificación se logra mediante una arquitectura jerárquica, que involucra convoluciones 3D y vectores densos embebidos que codifican la información del movimiento. Los resultados muestran que la arquitectura tiene una buena capacidad discriminativa y puede brindar indices de afectación del Parkinson. Además, la arquitectura brinda un conjunto de activaciones espacio-temporales que en las primeras capas puede describir las principales bases de representación del temblor. También, la obtención de mapas de explicabilidad obtenidos mediante la propagación de las probabilidades asociadas a cada población objeto de estudio, se pueden establecer como herramienta de apoyo permitiendo la visualización de las principales regiones asociadas por la red a la hora de la predicción.

1. LAS REPRESENTACIONES NEURONALES PROFUNDAS Y EL PARKINSON

En este capitulo se abordan los conceptos teóricos que fundamentan el trabajo de investigación desarrollado. Es por ello que se dará una exposición de las estrategias convolucionales como métodos del estado del arte para clasificación de patrones visuales. También se abordará una descripción del temblor y su estricta relación con el Parkinson.

1.1. REPRESENTACIONES CONVOLUCIONALES

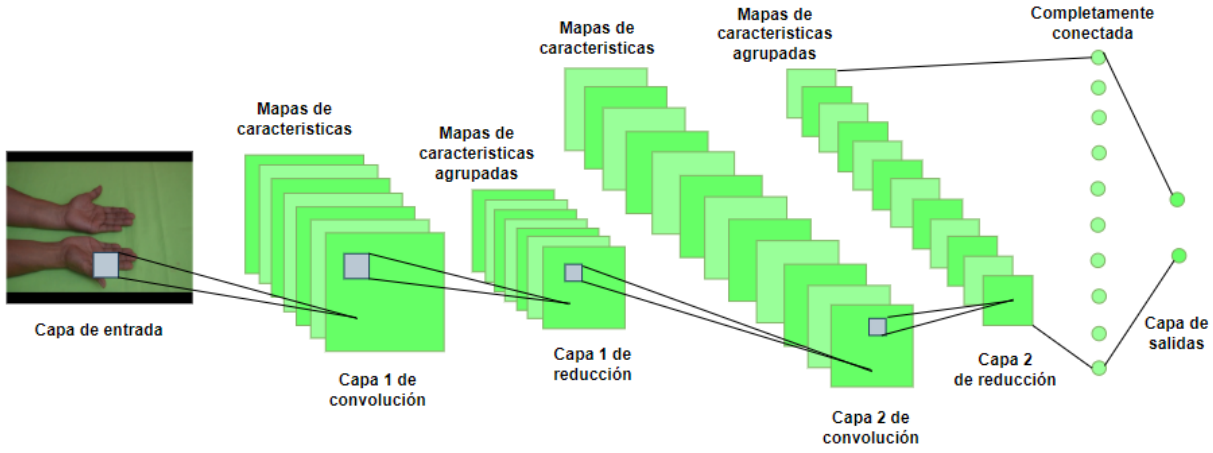
Actualmente, las redes neuronales convolucionales (CNN, por sus siglas en inglés) han demostrado relevancia en el campo de la visión por computador en tareas como clasificación de imágenes, reconocimiento facial, reconocimiento de acciones en video, detección de patrones, localización de objetos tanto en imágenes como en videos, entre otras [6]. Es por ello que en este trabajo se consideran estas arquitecturas como una potencial alternativa para la estimación, cuantificación y clasificación de temblores asociados con la enfermedad de Parkinson. Estas redes consisten en múltiples capas de filtros convolucionales (de allí su nombre) de una o más dimensiones. Después de cada capa se añade una función para realizar un mapeo causal no lineal. La fase de extracción de características se compone de capas alternas de neuronas convolucionales y etapas de reducción dimensional (etapa conocida como *pooling*). A medida que la arquitectura extrae patrones de alto nivel (mayor profundidad), la dimensionalidad de los datos se disminuye, siendo las neuronas en capas lejanas mucho menos sensibles a perturbaciones en los datos de entrada, pero al mismo tiempo, siendo estas activadas por características cada vez más complejas (con

---

[6] Yann LeCun, Yoshua Bengio y Geoffrey Hinton. "Deep learning". En: *nature* 521.7553 (2015), págs. 436-444.

un sentido semántico, de alto nivel, asociado a la tarea de aprendizaje) [7].

Figura 1. Tarea de clasificación tradicional por medio de una CNN donde para cada para una determinada entrada una serie de filtros de convolución extrae caracteristicas locales que en capas superiores se relacionan con la tarea de clasificación de interés.



En la Figura 1 se ilustra una arquitectura típica convolucional para resolver una tarea de clasificación. En este caso, los filtros convolucionales se especializan en recuperar patrones de las imágenes que puedan ser útiles para resolver una tarea específica de clasificación. La descomposición de la información de entrada en sus características y su relación local se aprende a través de una serie de capas que componen la red, logrando diferentes niveles de procesamiento y relación en cada nivel. En este sentido, representaciones convolucionales profundas describirán características de alto nivel con un sentido semántico, mientras que las primeras capas se especializarán en características primitivas como bordes y texturas. El nivel de detalle (profundidad de la red), conlleva el aprendizaje de una determinada cantidad de parámetros directamente relacionada a la profundidad. De esta manera, este nivel de detalle depende de la cantidad de datos disponibles para el entrenamiento de dichos parámetros otorgando a su vez una mayor validez y capacidad de generalizar el problema en términos de clasificación.

---

[7] Dan Claudiu Ciresan y col. "Flexible, high performance convolutional neural networks for image classification". En: *Twenty-second international joint conference on artificial intelligence*. 2011.

1.1.1. Representaciones profundas para análisis de video: Las arquitecturas convolucionales han sido extendidas a aplicaciones de análisis de video, capturando estructuras espacio-temporales desde bloques convolucionales volumétricos [8]. Estas representaciones, sin embargo, son típicamente aprendidas al nivel de unos pocos cuadros de video, que no logran modelar las acciones en toda su extensión temporal. A pesar de esto, las redes neuronales convolucionales a largo plazo (LTC, por sus siglas en inglés) han permitido aprender representaciones de video con extensiones temporales, mejorando así, el reconocimiento de acciones e incluyendo el modelamiento de largos desplazamientos capturados inclusive en baja resolución y con pocos cuadros de video.

Otras alternativas han sido expuestas en el estado del arte para el análisis continuo de secuencias de video. Una de ellas son las redes recurrentes (RNN, por sus siglas en inglés) que han permitido ser útiles para analizar series de tiempo que exhiben un comportamiento dinámico. A diferencia de otras representaciones profundas típicas, las redes recurrentes están dedicadas a simular unidades de memoria (capas profundas dedicadas a preservar la información histórica de los datos), manteniendo una relación de los eventos sucedidos en pasos de tiempo anteriores [9]. En las RNN, cada capa tiene una conexión desde las salidas del paso de tiempo anterior. La primera capa realiza una transformación lineal tanto para las entradas del paso del tiempo anterior como para las entradas de la capa actual. La suma de la información actual con la recibida del paso previo se da mediante funciones no lineales, generalmente sigmoidea, tangente hiperbólica (tanh) o unidad lineal rectificada (ReLU) [10]. En cuanto al análisis de video, se han propuesto alternativas híbridas que combinan operaciones convolucionales (representación espacial al nivel

---

[8] Gül Varol, Ivan Laptev y Cordelia Schmid. "Long-Term Temporal Convolutions for Action Recognition". En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6 (2018), págs. 1510-1517. DOI: 10.1109/TPAMI.2017.2712608.

[9] Neil J Vickers. "Animal communication: when i'm calling you, will you answer too?" En: *Current biology* 27.14 (2017), R713-R715.

[10] Zhenqi Xu, Jiani Hu y Weihong Deng. "Recurrent convolutional neural network for video classification". En: *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2016, págs. 1-6.

de la imagen) y unidades recurrentes (a nivel temporal) para tareas de clasificación. De esta manera, la extracción de características densas y locales de la imagen se relacionan con las temporales a través de los diferentes cuadros consecutivos [11]. Estas unidades son diseñadas en estrategias que requieren el modelamiento de intervalos con latencias temporales significativas y correspondencias no locales en el eje temporal [12]. Si bien es cierto que las arquitecturas RNN mantienen las relaciones temporales de las entradas en forma de secuencia, no explotan las características locales propias del dominio de las imágenes desaprovechando estas relaciones claves [10].

1.1.2. Explicabilidad e interpretablidad en representaciones profundas: A pesar de los significativos avances reportados por las arquitecturas CNN sobre diferentes dominios en tareas de clasificación, reconocimiento automático, entre otras; muchos de estos modelos reportan resultados destacables pero desarrollados sobre dominios y conjuntos de datos específicos, siendo difícil lograr su generalización y transferencia en otros contextos. El aprendizaje logrado a través de las múltiples capas que componen el modelo no permite explicar de manera clara y satisfactoria la decisión tomada por la arquitectura independientemente de la tarea objeto de estudio. Esta explicabilidad en las decisiones del modelo toma un rol importante en todos los dominios donde esta técnica incursione, resultando fundamental en el campo médico, donde se quiere ofrecer herramientas de apoyo clínico que respalden las decisiones, reduzcan la variabilidad del diagnóstico y sean soporte para tratamientos subsecuentes.

Aclarar e intentar interpretar las probabilidades de salida de las arquitecturas profundas es entonces un tema clave para interpretar y explicar una salida, asociada a una entrada específica. Esta explicación pretende entonces buscar la coherencia de lo predicho y así, dar confianza a

---

[11] Alex Graves. "Supervised sequence labelling". En: *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, págs. 5-13.

[12] Sepp Hochreiter. "Ja1 4 rgen schmidhuber (1997)."long short-term memory"". En: *Neural Computation* 9.8 ().

las predicciones, como también respaldar decisiones que resulten fácilmente justificables, de acuerdo al dominio de aplicación. De hecho, muchas de estas explicaciones e interpretaciones de los modelos que dan soporte en procesos implicados en decisiones, están siendo reglamentados y siendo legalmente obligatorias [13].

En busca de dar respuesta a estos cuestionamientos, recientemente, dentro del área de aprendizaje de máquina ha surgido toda una línea de trabajo que busca dar esta interpretabilidad a los modelos basados en datos. De esta manera, a partir de la estimación de probabilidad (el ¿qué?), se entabla la búsqueda que explique de manera suficiente el cómo se generó dicha probabilidad (el ¿por qué?) dando una mayor formalidad al problema [14]. Estos esquemas que se centran en buscar el por qué operan desde el vector de probabilidades resultante de la proyección de determinada entrada a lo largo de las diferentes capas que componen la arquitectura convolucional. De esta manera, la mayor probabilidad en el último vector es propagado a través de la red, multiplicando las activaciones para resaltar las regiones que tuvieron una mayor ponderación en esta decisión.

En el estado del arte existen diversidad de métodos que interpretan las tareas desempeñadas por los modelos, permitiendo de esta manera la estimación de aquellas zonas de interés para la red a la hora de tomar determinada decisión. Muchos de los estos métodos hacen uso del esquema completo del modelo actuando entre los datos de entrada, las predicciones de salida y aquellas activaciones en las capas ocultas que dieron lugar a dicha predicción. De esta manera, una visualización de las zonas relevantes sobre los datos de entrada permite la interpretación de los resultados dados por la red. A continuación una breve explicación de algunos de estos métodos:

---

[13] Jeya Vikranth Jeyakumar y col. "How can i explain this to you? an empirical study of deep neural network explanation methods". En: *Advances in Neural Information Processing Systems* 33 (2020), págs. 4211-4222.

[14] Finale Doshi-Velez y Been Kim. "Towards a rigorous science of interpretable machine learning". En: *arXiv preprint arXiv:1702.08608* (2017).

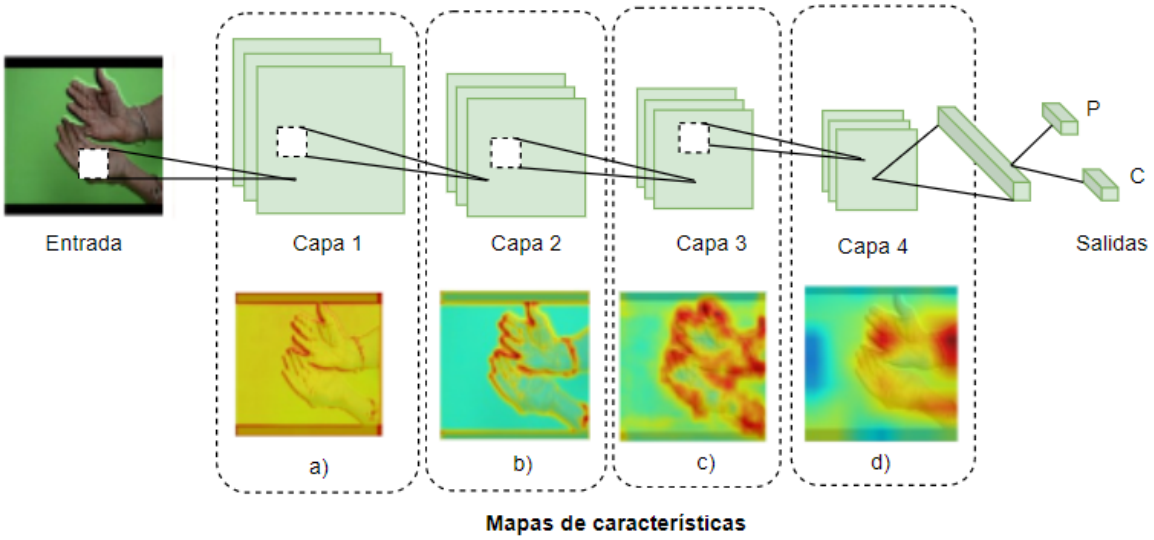
Mapas de características: Las CNN aprenden la correlación local de características y conceptos abstractos partir de filtros convolucionales. Este aprendizaje se da desde las características básicas (bordes y patrones texturales principalmente) desde las primeras capas, pasando por representaciones intermedias que toman por entrada estas primitivas para correlacionarlas en características más complejas (formas) hasta lograr su correspondencia con descriptores semánticos complejos (capas superiores).

En este sentido, una vez la arquitectura ha sido adaptada a un dominio de interés sobre una tarea particular, los filtros aprendidos generan mapas de activaciones que relacionan de manera visual los patrones locales propios de la imagen y permiten su visualización en los pixeles de las regiones con mayor relevancia para la red. Estas activaciones pueden ser extraídas en diferentes capas de la arquitectura para observar las principales activaciones de la imagen particular de entrada. Por ejemplo, como se puede evidenciar en la Figura 2 estas primeras capas aprenden características como bordes y texturas simples (columnas a) y b)), las capas posteriores aprenden características como texturas y patrones más complejos (columna c)) y finalmente las últimas capas aprenden características como objetos o localizaciones relacionadas con los objectos que principalmente aportan a la tarea de clasificación (columna d)).

Entonces, mapas de interpretabilidad pueden obtenerse a partir del bloque de activaciones en una instancia particular de la representación. Para ello, algunos trabajos en la literatura han realizado *pooling* de mapas en un mismo bloque, condensando las activaciones localizadas más relevantes, convirtiéndose así en esquemas que dan información sobre la información saliente, tenida en cuenta para una probabilidad particular. Por ejemplo, haciendo *average-pooling* se hace un promedio a través de las activaciones para buscar los valores medios en el bloque de activaciones. También operadores no-lineales como el *max-pooling* permiten condensar las activaciones más importantes a través del bloque.

Los métodos CAM y Grad-CAM (*Class Activation Maps*): Una de las estrategias más conocidas en la literatura para generar mapas de explicabilidad ante una predicción de una red particular, son los métodos basadas en los mapas de activación por clase (CAM, por sus

Figura 2. Visualización de mapas características para una muestra de un paciente con parkinson en configuración postural. En la primera fila se observa la arquitectura de la red a partir de la entrada. En la segunda fila se observa la visualización para las capas convolucionales 1, 2, 3 y 4 respectivamente. Se evidencia que las primeras capas se fijan en bordes y texturas simples.



siglas en inglés) [15]. En este sentido, el método más clásico CAM permite visualizar las regiones de la imagen que tienen mayor aporte respecto a una predicción particular de clase. En este caso, existe un proceso de *Global Average Pooling (GAP)* entre la última capa convolucional y la última capa *full conected*, relacionada con la predicción. Por un lado, la capa convolucional tiene una representación que mantiene relaciones espaciales del objeto y las activaciones dispuestas para una entrada particular. Por otro lado, la capa densa describe las iteraciones complejas entre la imagen, resultando en un vector embebido que conduce a la clasificación.

Entonces el proceso de GAP permite obtener un único mapa de activaciones, procesado como el promedio de las activaciones, a través del eje de los filtros. Este mapa es entonces ponderado con los respectivos pesos que conectan la capa densa, como: $M_c(x, y) = \sum_k w_k^c f_k(x, y)$ donde w_k^c son los pesos que permiten promediar espacialmente los gradientes que fluyen hacia atrás desde la

[15] Bolei Zhou y col. "Learning deep features for discriminative localization". En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, págs. 2921-2929.

probabilidad de salida. Por otra parte, $f_k(x, y)$ corresponde al *k-esimo* mapa de características. Estos pesos w_k^c se interpretan como la relevancia de cada mapa de características $M_c(x, y)$ hacia una clase c determinada permitiendo visualizar el proceso de toma de decisiones ejercido por la red neuronal. Estos mapas CAM originales han tenido gran relevancia como herramienta de soporte en la interpretabilidad, pero resultan limitados al cálculo en la última capa convolucional y tienen restricciones en cuanto al resumen del mapa de características.

Recientemente, para suplir estas restricciones, se propuso una mejora a esta estrategia, a través de los mapas Grad-CAM [16]. Está estrategia es flexible ya que permite ser operada y extendida a cualquier arquitectura convolucional, donde se calculan los gradientes que fluyen hacia atrás con respecto a una determinada clase c y se multiplican por una capa convolucional seleccionada. Es decir, este método usa la información del gradiente que fluye desde una determinada clase de salida hacia la última capa convolucional de la CNN para medir la importancia de la decisión tomada. Similar a CAM, los mapas de calor de Grad-CAM son la combinación ponderada de los mapas de características seguidas por una función no lineal ReLU. $M_c(x, y) = ReLU(\sum_k w_k^c f_k(x, y))$ este peso w_k^c representa una linealización parcial de la red profunda para cierta clase c.

1.2. TEMBLOR EN PARKINSON

La enfermedad de Parkinson (EP) es un síndrome neurodegenerativo que provoca síntomas como: temblor, rigidez, bradicinesia (lentitud en el movimiento) e inestabilidad postural [17]. En aproximadamente la mitad de los pacientes, el síntoma más visible es el temblor y únicamente el 10 % de los pacientes diagnosticados con la enfermedad no lo presentan [5]. Los temblores varían en frecuencia y amplitud, donde su categorización se basa según la posición o postura ejerciendo

---

[16] Ramprasaath R Selvaraju y col. "Grad-cam: Visual explanations from deep networks via gradient-based localization". En: *Proceedings of the IEEE international conference on computer vision*. 2017, págs. 618-626.

[17] Werner Poewe y col. "Parkinson disease". En: *Nature reviews Disease primers* 3.1 (2017), págs. 1-21.

el movimiento necesario para provocarlo [18]. Por ejemplo, cuando no se ejerce ninguna influencia para observar el temblor, este se denomina en reposo. También existe la categoría de temblor postural donde se solicita al paciente levantar las manos, para que la acción de la gravedad actúe como una fuerza externa que permita amplificar las oscilaciones del temblor y tener una caracterización de esta, en estadios tempranos o sospechosos. Lo anterior se puede visualizar en la Figura 3.

El temblor en Parkinson se describe clásicamente como temblor de reposo, sin embargo, Koller WC *et. al.*, en [19] documentó que el 92 % de una serie de 50 pacientes presentó temblor postural, existiendo una combinación de temblor postural y de reposo en 76 % de ellos. Se presentó incluso que en pacientes con EP y temblor leve puede coexistir el temblor en reposo con temblor postural. El temblor en reposo se presenta con una amplitud y una frecuencia modal que varía entre 4 y 6 Hz, mientras que en el temblor postural varía entre 5 y 12 Hz, frecuencia que incrementa con estrés o ansiedad y disminuye con movimientos voluntarios [20].

Particularmente, el temblor de manos en reposo es uno de los biomarcadores más importantes en la EP, este indicador se describe como un movimiento periódico y oscilatorio cuando las manos se encuentran apoyadas, es decir, sin contracción muscular voluntaria. En el análisis clínico estándar, dicha caracterización es difícil de observar debido a la baja amplitud del movimiento especialmente en las primeras etapas de la enfermedad [21]. En la práctica se lleva a cabo una exageración física del temblor con el fin de destacar las manifestaciones perceptuales y de solventar dicha problemática, esta exageración se conoce como configuración postural. En

---

[18] Ahmad Anouti y William C Koller. "Tremor disorders. Diagnosis and management." En: *Western journal of medicine* 162.6 (1995), pág. 510.

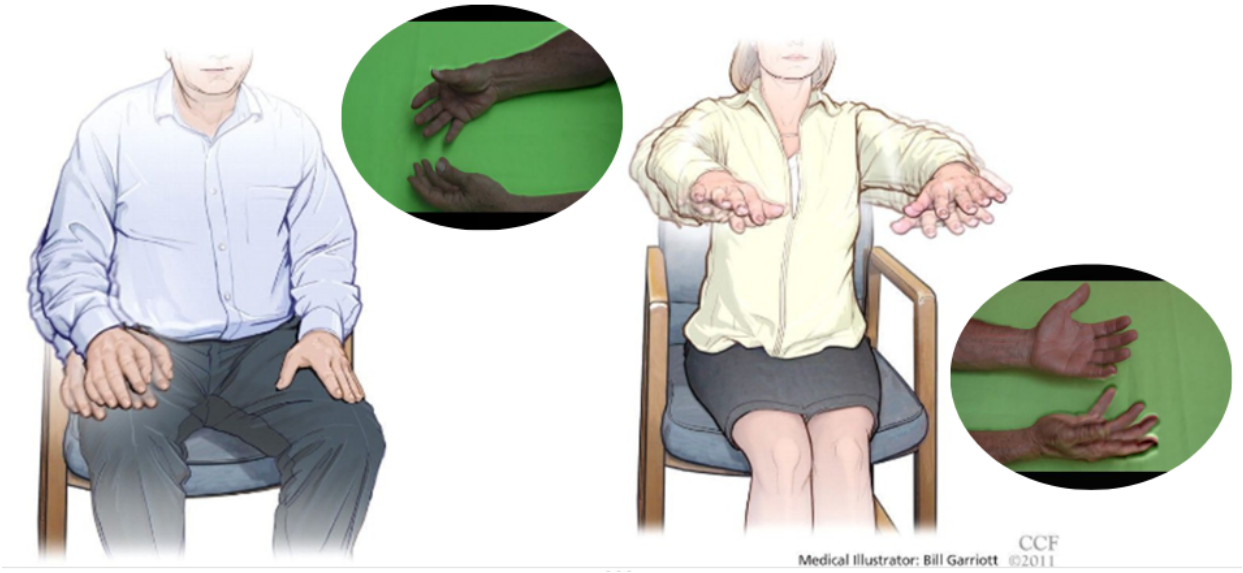
[19] William C Koller, Bridget Vetere-Overfield y Ruth Barter. "Tremors in early Parkinson's disease." En: *Clinical neuropharmacology* 12.4 (1989), págs. 293-297.

[20] David E Vaillancourt y Karl M Newell. "The dynamics of resting and postural tremor in Parkinson's disease". En: *Clinical Neurophysiology* 111.11 (2000), págs. 2046-2056.

[21] Jacopo Pasquini y col. "Progression of tremor in early stages of Parkinson's disease: a clinical and neuroimaging study". En: *Brain* 141.3 (2018), págs. 811-821.

esta configuración, las manos permanecen sin ningún soporte y por tanto, la fuerza de gravedad actúa como una sobrecarga para aumentar el movimiento de las manos [22]. Sin embargo, la exageración del temblor postural produce contracciones musculares voluntarias consideradas como señales ruidosas con respecto al temblor parkinsoniano.

Figura 3. Representación de las configuraciones para evaluar el temblor. A la izquierda temblor en reposo en donde las manos se encuentran apoyadas. A la derecha temblor postural en donde las manos se mantienen suspendidas, donde la gravedad ejerce una fuerza en los brazos. Se observa que en la elipse del lado derecho la manos se encuentran apoyadas sobre una mesa. En la elipse del lado izquiero las manos se encuentran suspendidas. Ilustración tomada y adaptada de: [23]



1.3. CUANTIFICACIÓN Y SOPORTE DE PATRONES DE TEMBLOR

Usualmente, el monitoreo y el diagnóstico de la EP está soportado mediante mediciones cuantitativas tomadas por medio de acelerómetros y giroscopios, capturando de esta manera diferentes

---

[22] Jie Zhang y col. "Differential diagnosis of Parkinson disease, essential tremor, and enhanced physiological tremor with the tremor analysis of EMG". En: *Parkinson's Disease* 2017 (2017).

direcciones de aceleración y variaciones angulares asociadas con el temblor[24]. Por ejemplo, sistemas basados en sensores inerciales colocados en las manos y brazos calculan cambios de velocidad y aceleración en configuraciones de descanso y postural. Así, se identifica que las medidas de las manos son más significativas que las medidas del antebrazo para la caracterización del temblor[24]. Análogamente, con los sensores inerciales son discriminadas las dimensiones cinemáticas entre sujetos control y pacientes con Parkinson [25]. Sin embargo, estos métodos cinemáticos tienen limitaciones relacionadas a la cuantificación de movimientos sutiles asociados a etapas iniciales. Alternativamente, técnicas basadas en visión por computadora para la magnificación de movimiento han avanzado en el análisis de desplazamientos sútiles en el video. Por ejemplo, en la literatura se reporta el enfoque de magnificación de video de tipo Eureliano. Este enfoque asume que el objeto de interés permanece relativamente estático a la cámara y exagera los movimientos sútiles de la grabación. En este caso, se hace una descomposicion temporal del video utilizando una pirámide laplaciana. A partir de esta descomposición se seleccionan las frecuencias temporales de interés las cuales son amplificadas. Finalmente, las diferentes bandas de frecuencia temporal se agrupan y se hace una reconstrucción del video exaltandose las frecuencias de interés del video que tipicamente corresponde a los movimientos sútiles del video. Esto puede ser una herramienta de soporte para el especialista con el fin de identificar las partes del cuerpo con sus diferentes niveles de afectación. A pesar de esto, esta estrategia depende de la fijación apropiada de filtros temporales, que permitan recuperar señales relacionadas con el temblor y aislarlas de otros posibles movimientos registrados en el video. A continuación se detallan las diferentes propuestas que existen en el estado del arte para cuantificar el temblor, según los dispositivos.

---

[24] Adriano de Oliveira Andrade y col. "Task-Specific Tremor Quantification in a Clinical Setting for Parkinson's Disease". En: *Journal of Medical and Biological Engineering* 40.6 (2020), págs. 821-850.

[25] Weiguang Huo y col. "A heterogeneous sensing suite for multisymptom quantification of Parkinson's disease". En: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28.6 (2020), págs. 1397-1406.

1.3.1. Estrategias clásicas y/o invasivas: Actualmente, la evaluación clínica del temblor se logra mediante escalas como la Escala de Medición de la Enfermedad de Parkinson Unificada (UPDRS) que se basa en patrones identificados a través de la observación [26]. Sin embargo, estas observaciones están sujetas a la experiencia de quien realiza las valoraciones proporcionando una evaluación subjetiva y de gran variabilidad [27]. Por lo tanto, se han propuesto varios enfoques para la evaluación computacional del temblor, lo cual permite reducir dicha subjetividad en el análisis.

Timmer *et. al.* en [28] en su trabajo extrajo características cuantitativas de la aceleración del temblor en configuración postural a través del tiempo, con el fin de separar tres tipos de temblores (fisiológico, esencial y parkinsoniano). Otros métodos como el análisis espectral también han abordado la detección y cuantificación del temblor [29][30]. Un método computacional relacionado con el análisis espectral para la cuantificación del temblor se presenta por Riviere *et. al.* en [29], determinando la frecuencia y la amplitud del temblor ajustando sus pesos mediante *Weighted Frequency Fourier Linear Combiner (WFLC)*, la ponderación de los coeficientes de Fourier. Al cuantificar las características que varían en el tiempo, el WFLC ayuda a interpretar correctamente los resultados del análisis espectral proporcionando una representación más precisa del temblor.

---

[26] Sandrine Greffard y col. "Motor score of the Unified Parkinson Disease Rating Scale as a good predictor of Lewy body–associated neuronal loss in the substantia nigra". En: *Archives of neurology* 63.4 (2006), págs. 584-588.

[27] Carmen Rodriguez-Blazquez, Maria João Forjaz y Pablo Martinez-Martin. "Rating scales in movement disorders". En: *Movement Disorders Curricula*. Springer, 2017, págs. 65-75.

[28] Jens Timmer y col. "Characteristics of hand tremor time series". En: *Biological cybernetics* 70.1 (1993), págs. 75-80.

[29] Cameron N Riviere, Stephen G Reich y Nitish V Thakor. "Adaptive Fourier modeling for quantification of tremor". En: *Journal of neuroscience methods* 74.1 (1997), págs. 77-87.

[30] Malenka Mader y col. "Spectral and higher-order-spectral analysis of tremor time series". En: *Clin Exp Pharmacol* 4.149 (2014), págs. 2161-1459.

Por otro lado, gran parte del trabajo reciente se basa en el uso de acelerómetros [31][32] y electromiografía (EMG) para el monitoreo a largo plazo de los pacientes. Por ejemplo, Bacher *et. al.* presentó un método que permitió la cuantificación continua del temblor parkinsoniano, utilizando EMG, durante un periodo de hasta 24 horas [33]. En este trabajo, se calculó la ocurrencia, intensidad y frecuencia del temblor. Este trabajo presenta una ventaja considerable dado que la cuantificación a largo plazo proporciona datos confiables para estudiar las variaciones del temblor y controlar los efectos del tratamiento médico. No obstante, este sistema puede considerarse invasivo al implicar sensores y un posicionamiento de los mismos en diferentes partes del cuerpo para lograr la detección del movimiento afectando la naturalidad del mismo.

1.3.2. Estrategias basadas en cuantificación de video: Diferentes trabajos se han dedicado al desarrollo de estrategias basadas en cuantificación de características de video para la estimación de temblores, teniendo en cuenta su marcada ventaja para el monitoreo no invasivo. Como muestra de ello, Uhríková *et. al.* analizó la frecuencia del temblor de manos a partir de secuencias de video. Para tal fin usó el cambio de intensidad del área local a tiempo [34]. Los resultados de este trabajo se compararon con las medidas obtenidas por medio de un acelerómetro. Otra propuesta, la reportó Soran *et. al.* donde clasificó la presencia o ausencia de temblor en

---

[31] Shyamal Patel y col. "Using wearable sensors to predict the severity of symptoms and motor complications in late stage Parkinson's Disease". En: *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2008, págs. 3686-3689.

[32] Robert LeMoyne, Cristian Coroian y Timothy Mastroianni. "Quantification of Parkinson's disease characteristics using wireless accelerometers". En: *2009 ICME International Conference on Complex Medical Engineering*. IEEE. 2009, págs. 1-5.

[33] M Bacher, E Scholz y HC Diener. "24 hour continuous tremor quantification based on EMG recording". En: *Electroencephalography and clinical neurophysiology* 72.2 (1989), págs. 176-183.

[34] Zdenka Uhríková y col. "Action tremor analysis from ordinary video sequence". En: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2009, págs. 6123-6126.

videos mediante *support vector machine* (SVM), con *kernel* radial [35]. En este trabajo se calculó el flujo óptico de las manos y los coeficientes de la transformada discreta del coseno para extraer las frecuencias y los cambios de direcciones del movimiento.

Sin embargo, la estrategia tiene limitaciones para detectar temblores de amplitud leve, debido a que el flujo óptico es un método propuesto para cuantificar objetos en movimiento y no objetos en donde el movimiento es apenas perceptible. Una solución en la identificación de los temblores leves es el enfoque propuesto por Contreras *et. al.* [36], el cual se basa en la magnificación del movimiento en el rango de frecuencias entre 4 y 6 Hz mediante el enfoque euleriano. Posteriormente, se propone la cuantificación del temblor como la varianza de la intensidad de los pixeles en puntos seleccionados a conveniencia (en los dedos). Este trabajo magnifica movimientos sútiles, sin embargo; los puntos de selección de interés son seleccionados manualmente. Específicamente en este trabajo se marcan los puntos de interés que corresponden a los extremos de los dedos que pueden generar un gradiente visible con respecto al fondo. Estos puntos, sin embargo, pueden variar con respecto a cada notación sesgando el anális por la subjetividad de la anotación. Adicionalmente, la reducción del movimiento a puntos fijos simplifica la información del movimiento considerandose una potencial pérdida significativa del mismo.

1.3.3. Estrategias basadas en aprendizaje profundo: El aprendizaje profundo permite aprender representaciones de datos con múltiples niveles de abstracción. Muchos estudios han demostrado su efectividad cuando se cuenta con grandes volúmenes de datos [6]. Respecto al tema objeto de estudio, Kim Han *et. al.* propuso un sistema de evaluación del temblor basado en CNN para diferenciar la gravedad de los síntomas medidos en datos recopilados a través de un dispositivo portátil equipado con un acelerómetro y un giroscopio montados en un módulo de

---

[35] Bilge Soran y col. "Tremor detection using motion filtering and SVM". En: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE. 2012, págs. 178-181.

[36] Sergio Contreras, Isail Salazar y Fabio Martínez. "Parkinsonian hand tremor characterization from magnified video sequences". En: *14th International Symposium on Medical Information Processing and Analysis*. Vol. 10975. International Society for Optics y Photonics. 2018, pág. 1097503.

la muñeca [37]. Los datos medidos se transformaron en el dominio de frecuencia y se usaron para construir una imagen bidimensional para entrenar la CNN. Otro trabajo presentado por Zheng *et. al.* demostró la viabilidad de utilizar los datos de aceleración, respaldados por algoritmos de aprendizaje profundo para cuantificar la severidad del temblor [38]. En este trabajo se utilizó un sistema de variables de aceleración en tres direcciones ubicados en los brazos de 20 personas en estudio. Mientras la persona desempeñaba las tareas de: beber, extender el brazo, tocarse la nariz, poner un vaso sobre otro, dibujar y escribir; las señales eran capturadas, filtradas y finalmente clasificadas. Se crearon modelos de clasificación de actividad (ACM) y modelos de evaluación del temblor (TEM) con la implementación de algoritmos que distinguen las actividades humanas voluntarias y evaluar la gravedad del temblor, respectivamente. A pesar de que las arquitecturas propuestas en estos trabajos reportan una alta precisión de estimación en los resultados, sus desarrollos siguen basados en el uso de dispositivos en el que los sensores juegan un papel importante a la hora de realizar mediciones, generando inconvenientes debido a la sensibilidad de los mismos.

---

[37] Han Byul Kim y col. "Wrist sensor-based tremor severity quantification in Parkinson's disease using convolutional neural network". En: *Computers in biology and medicine* 95 (2018), págs. 140-146.

[38] Xiaochen Zheng y col. "Activity-aware essential tremor evaluation using deep learning method based on acceleration data". En: *Parkinsonism & related disorders* 58 (2019), págs. 17-22.

2. PLANTEAMIENTO Y JUSTIFICACIÓN DEL PROBLEMA

Durante la rutina clínica, las primeras valoraciones médicas que se realizan a los pacientes con síntomas que afectan al movimiento, como el temblor de manos, se evalúan de manera observacional, siguiendo esquemas posturales y de reposo en donde los médicos intentan determinar si los patrones anormales se encuentran asociados a la EP. En estadios tempranos de la enfermedad esta tarea presenta complicaciones puesto que en las primeras etapas los movimientos anormales en posición de reposo se presentan en magnitudes muy bajas haciéndolos casi imperceptibles a la vista humana. A pesar de que en la configuración postural se logra una exageración del patrón del temblor al suspender las extremidades, haciendo más perceptibles las observaciones, estas pueden ser afectadas por contracciones musculares voluntarias consideradas como señales ruidosas. Adicional a lo anterior, estas valoraciones presentan también una variabilidad subjetiva, ya que están sujetas al criterio y a la experticia de los médicos generando problemas de precisión al momento de dar resultados orientados al diagnóstico, al seguimiento y al control de la enfermedad, debido a que dependen totalmente de una percepción visual.

Por otro lado, considerando que existen herramientas computacionales que permiten cuantificar los patrones anormales del movimiento, estas en su mayoría se basan en técnicas invasivas en donde se requiere el uso de dispositivos electrónicos y/o sistemas portátiles basados en sensores para obtener mediciones y hacer un correcto análisis de los datos. También existen enfoques de estrategias basadas en aprendizaje profundo, pero algunas siguen siendo invasivas al implicar selección de sensores dado a que generan una alteración en los movimientos naturales y la posible distorsión o pérdida de información debido a la sensibilidad de los mismos. Existen otros trabajos que han realizado análisis de temblor logrando cuantificar la EP utilizando videos. Por ejemplo, realizando magnificación de las secuencias de video y luego realizando una marcacion de puntos de interés para analizar su variacion temporal, y su posterior asociación con el temblor. Sin embargo, esta intervención manual puede ser subjetiva y restrictiva para el análisis del mismo.

Además la selección de puntos de interés es escasa, lo cual limita la caracterización del temblor. Además, la literatura carece de representaciones visuales que permitan explicar los patrones relevantes identificados a partir de los videos, siendo esto un campo inicial de investigación.

3. OBJETIVOS

OBJETIVO GENERAL

Desarrollar una estrategia de aprendizaje profundo para la clasificación y representación visual del temblor de manos asociados a pacientes con Parkinson.

OBJETIVOS ESPECIFICOS

- Procesar un conjunto de videos de temblor de manos de pacientes con Parkinson y pacientes control en configuraciones de reposo y postural.
- Implementar una estrategia de aprendizaje profundo que permita clasificar pacientes con patrones de movimiento asociados a la Enfermedad de Parkinson.
- Cuantificar mapas de representación visual del temblor de manos en pacientes con Parkinson.
- Validar el modelo propuesto en términos de capacidad de clasificar patrones de temblor asociados al Parkinson con respecto a patrones Control.

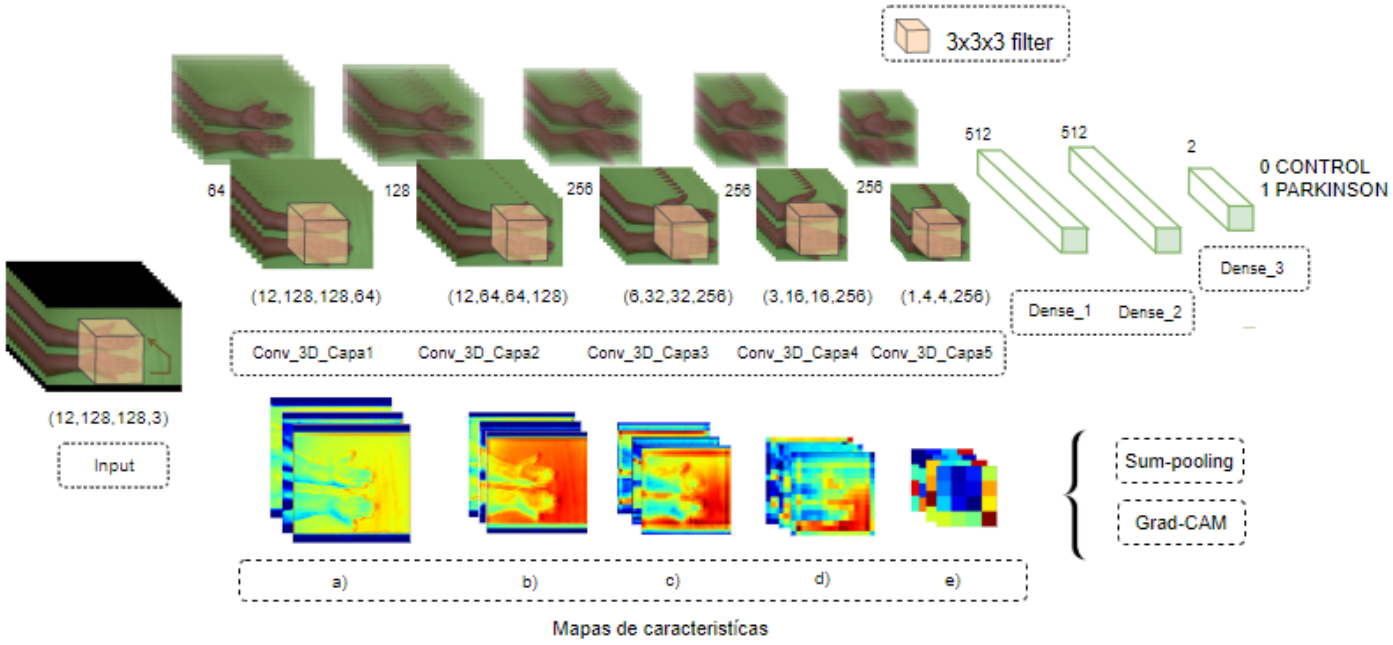
4. ENFOQUE PROPUESTO

Los patrones de temblor de manos varían en un rango entre 4 y 12 Hz y se pueden modelar como una combinación de temblor en reposo y temblor postural. El temblor de reposo está ausente en las actividades cotidianas, aumenta con el estrés y la ansiedad donde disminuye con los movimientos voluntarios. En el caso de los pacientes con EP, estos patrones de locomoción pueden cambiar su periodicidad dependiendo del sujeto y los marcadores temporales clásicos del temblor pueden describirse en secuencias de video más largas y variables. En este trabajo, utilizamos una estrategia de convolución 3D para resaltar y clasificar el temblor de manos parkinsoniano en secuencias de video de manera automática. También, como aplicación clínica, resulta relevante la generación de mapas explicativos que soporten las decisiones tomadas por la representación profunda entrenada. En este sentido, el presente trabajo plantea dos versiones de mapas de atención que permiten determinar de manera observacional las principales regiones en el video implicadas en cada predicción. Estos mapas de atencion de las activaciones y el Grad-CAM corresponden a representaciones visuales del temblor. Antes de hacer la proyección a la red convolucional se realizó el procesamiento de los datos que consistio en recortar los videos así como fijar los periodos temporales y también se válido con versiones magnificadas de los videos. Para la magnificación de los videos se utilizó una estrategia Euleriana, la cual consiste en amplificar un conjunto de frecuencias temporales de video a través de una proyección multiescala. En la Figura 4 se ilustra la metodología propuesta. A continuación se describen cada uno de sus componentes.

4.1. REPRESENTACIÓN VOLUMÉTRICA CONVOLUCIONAL

En este trabajo es de gran interés el estudio y codificación de patrones espacio-temporales registrados en secuencias de video. De hecho, el temblor puede ser entendido como un patrón rítmico localizado que puede ser codificado en ventanas temporales a través de video. En este sentido,

Figura 4. Representación de la metodología propuesta. En la parte superior, arquitectura convolucional 3D en donde cada fila representa diferentes entradas en la red convolucional. En la parte inferior mapas de características del modelo, siendo desde la a) hasta la e) los mapas de representación visual del temblor. Al lado derecho de los mapas de caracteristícas se evidencian las dos estrategias de cuantificación.



el modelamiento de patrones localizados, espacio-temporales del temblor, son logrados con representaciones de aprendizaje cuyas primeras capas de procesamiento involucran convoluciones 3D. Con esto en mente, los *kernels* volumétricos permiten el procesamiento y la cuantificación del temblor a lo largo del video. Estos filtros resultan ser los parámetros que la arquitectura necesita ajustar para realizar una adecuada extracción de características. Particularmente, sea el volumen de imágenes $\mathbf{I}(\mathbf{x})_t$ de dimensión $L \times H \times W$ que representa alguna secuencia de video extraída previamente donde L indica el número de cuadros temporales y H, W las dimensiones espaciales de los cuadros. Además, sea κ el *kernel* de convolución de dimensión (z, v, w), donde la dimensión z convoluciona sobre el eje temporal y las dimensiones v, w sobre los ejes espaciales. Con las restricciones $z \leq L$, $v \leq H$ y $w \leq W$, la salida al aplicar la operación de convolución de ψ sobre $\mathbf{I}(\mathbf{x})_t$ es $L' \times H' \times W' \times D'$, donde L', H' y W' representan las dimensiones reducidas del

volumen original y $|\Psi|$ el número de *kernels* usados. Es decir $\Phi(\mathbf{x}) = \sum_{g=1..q} \mathbf{I}(\mathbf{x})_t * \Psi_g$, donde $|\Phi|$ son los mapas de características espacio temporales diferentes, logrando capturar y caracterizar el movimiento. El *kernel* recorre el volumen de entrada en las tres direcciones teniendo en cuenta que cada posición volumétrica se obtendrá mediante una multiplicación y suma de los elementos generando una salida tridimensional rica en información espacio temporal. El mapa de salida de características b con valores en cada posición (t, y, x) para el j-ésimo *kernel*. El volumen total de características es la unión de cada mapa b para el numero Ψ de *kernels* especificados. Esta operación es el núcleo en una capa de red donde su generalización permite el uso de sucesivas capas de aprendizaje con el ajuste de más *kernels* de manera independiente.

4.2. ARQUITECTURA PROFUNDA CONVOLUCIONAL 3D

La arquitectura adoptada en este trabajo considera convoluciones espaciales y temporales con el objetivo de aprender las características espacio-temporales de manera local relacionando el temblor con en la EP mediante secuencias de video permitiendo su procesamiento de forma natural capturando los patrones propios de las configuraciones postural y de reposo. Mediante esta configuración arquitectural, se asume como información relevante aquella presente a través del tiempo y no en el espacio debido a que la red a medida que se hace profunda sacrifica la información espacial a costa de ganar información temporal. El presente trabajo propone una estrategia basada en una red convolucional 3D profunda que extrae el movimiento tanto a nivel espacial como temporal permitiendo obtener en cada capa de convolución un volumen de mapas de características interpretado como una firma asociada a cada población objeto de estudio permitiendo realizar una predicción que soporte el diagnóstico para la EP.

En esta arquitectura para cada capa L, las transformaciones lineales se calculan progresivamente, seguidas de no linealidades contractivas que proyectan la información en un conjunto de q filtros aprendidos expresados como: $\Psi_L = \{\Psi\}_{i=1}^{q}$. Así, $\mathbf{I}(\mathbf{x})_t$ es filtrada en la primera capa por Ψ_1. Obteniendo una representación: $\Phi_1(\mathbf{x}) = \sum_{g=1..q} \mathbf{I}(\mathbf{x})_t * \Psi_g$ donde Ψ_g siendo cada filtro de convolución aprendido de forma independiente. La representación resultante se convoluciona

sucesivamente con los filtros de las siguientes $N-1$ capas $\{\Psi_i\}_{i=2}^{N}$. La representación convolucional $\Phi_1(\mathbf{x})$ viene dada por las representaciones de cada capa:$\{\Phi_i(\mathbf{x})\}_{i=2}^{N}$. Las respuestas de las primeras capas de esta red generan características generalizadas y de bajo nivel, como bordes y texturas. Estas características de bajo nivel se combinan para construir representaciones más complejas en las capas convolucionales posteriores. Finalmente, las características que se encuentran en la última capa convolucional corresponden a características de nivel superior donde los datos de entrada logran una relación semántica con la etiqueta de predicción. La entrada de esta red convolucional son secuencias de video de temblor de manos, las cuales en capas sucesivas son densamente correlacionadas permitiendo un modelamiento robusto de los patrones de movimiento. En la figura 4 se ilustra el esquema convolucional propuesto para el modelamiento volumétrico de patrones espacio temporales con respecto a la capacidad de clasificar entre pacientes con Parkinson y pacientes control.

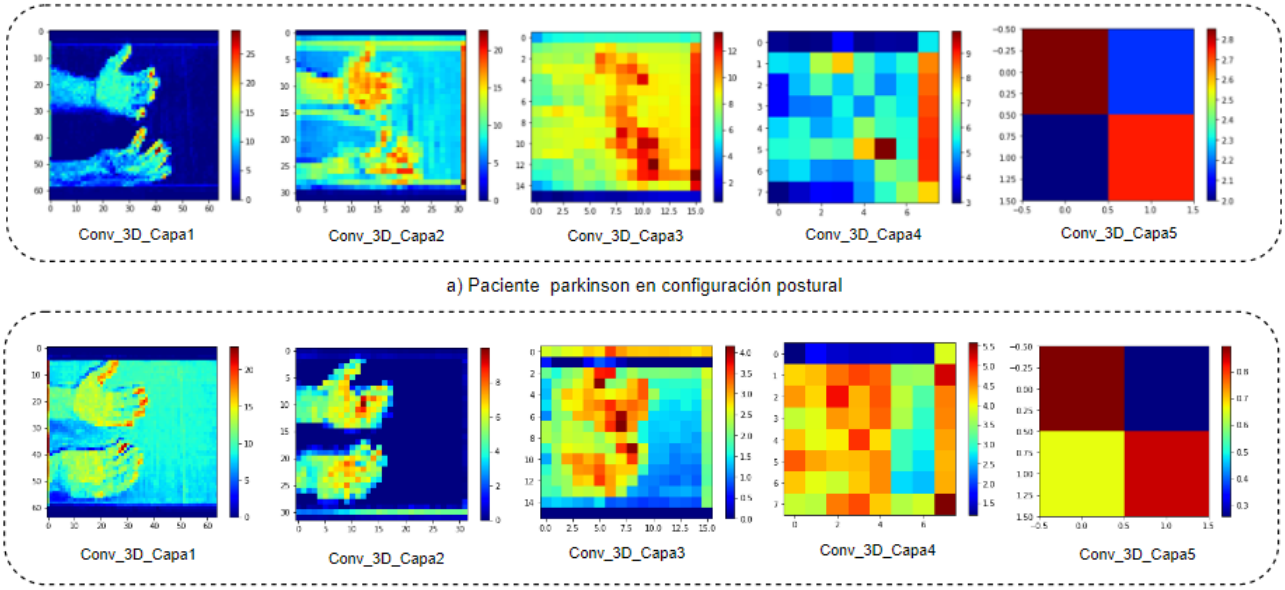
4.3. MAPAS DE EXPLICABILIDAD

La configuración de las CNN permite la representación jerárquica de las regiones locales de cierta entrada. En el caso particular de las configuraciones 3D, estas representaciones son enriquecidas por las relaciones entre cuadros sucesivos resultando relevantes en problemas relacionados a patrones observables a través de secuencias de video. Estos resultados, sin embargo, pueden estar sesgados por artefactos, condiciones de la captura, o correlaciones ocultas que pueden limitar la explicabilidad de los resultados obtenidos. Es por ello, que la capacidad de predicción por sí misma en un ambiente clínico resulta poco atractiva en términos de usabilidad por parte de expertos debido a la poca interpretación de las decisiones tomadas.

Con esto presente, el presente trabajo presenta ventajas explicacionales tanto en activaciones (respuestas de las convoluciones a una entrada particular), así como en la implementación de algoritmos que permitan vislumbrar el por qué de las predicciones realizadas por la red. En las siguientes subsecciones se explican las dos estrategias implementadas como respuesta a las necesidades de complementar las predicciones realizadas por la red mediante mapas que

permitan interpretar las regiones más comprometidas a la hora de realizar la predicción.

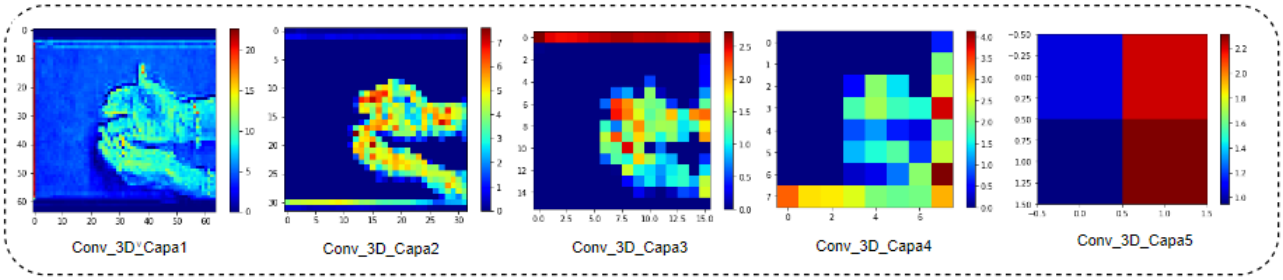
Figura 5. Representación de los mapas de características en las capas convolucionales. La fila superior evidencia los mapas de características asociados a un paciente con parkinson, mientras que la fila inferior para un paciente control.



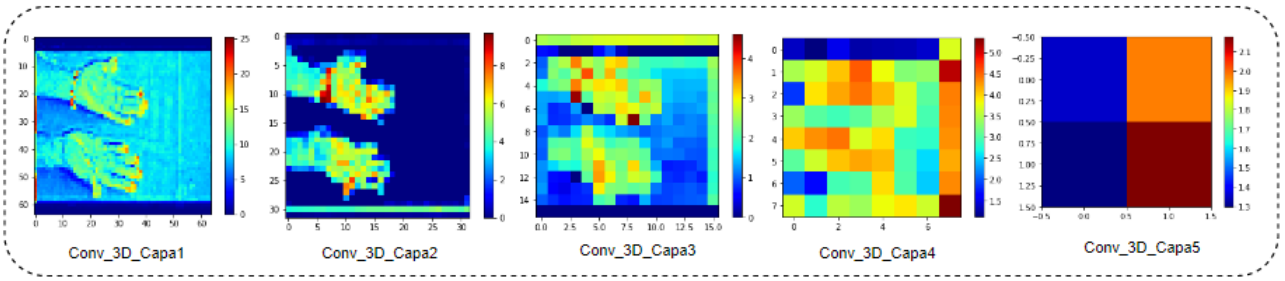
4.3.1. Agrupación y síntesis usando sumas entre filtros (sum-pooling): Una primera alternativa para generar mecanismos explicativos visuales son los mapas de activaciones generadas en cada capa convolucional. En este sentido, estas activaciones pueden ser relevantes en el análisis de comportamientos de temblor y ser tratados como una herramienta observacional que proporcione los principales segmentos espacio-temporales que se activan para cada una de las secuencias. En este caso, la estrategia podría ser utilizada como soporte durante el diagnóstico visualizando diferentes mapas de activación frente a diversas secuencias de entrada y a diferentes niveles de representación.

En el presente trabajo, estos mapas de características se obtienen a partir de secuencias de video capturadas sobre las manos en la configuración tanto postural como en reposo con el fin de determinar el temblor asociado a cada escenario y condición del paciente objeto de estudio.

Figura 6. Representación de los mapas de características en las capas convolucionales. La configuración de reposo de un paciente con parkinson y un paciente control se evidencian en las filas superior e inferior, respectivamente.



a) Paciente parkinson con configuración de reposo



a) Paciente control con configuración de reposo

En cada capa convolucional 3D, el video se proyecta a un conjunto de *kernels*, generando activaciones que están correlacionadas con el temblor o los patrones espacio-temporales que tuvieron mayor aporte en una predicción particular.

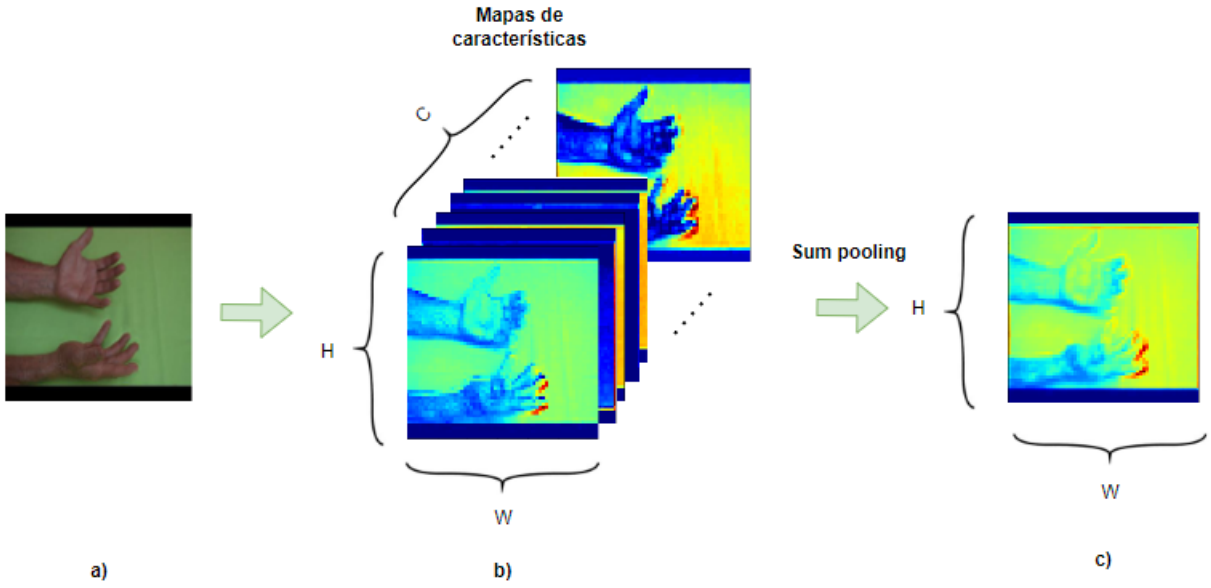
Las Figuras 5 y 6 evidencian la evolución jerárquica de los mapas de características a través de las capas que componen la arquitectura para pacientes Parkinson y control. Como se puede observar, las primeras capas centran su atención principalmente en la zona de los dedos sugiriendo ser un área clave para el diagnóstico donde una mayor correlación local de la velocidad puede explicar esta fijación por parte de la red.

En las capas sucesivas la atención no solo se centra en los dedos sino en las zonas donde se encuentran las manos, donde el fondo resulta ser descartado por la red debido a su influencia nula en el proceso de predicción. Como es de esperarse, las activaciones primarias resaltan patrones localmente correlaciones y se puede obtener una visualización consistente del temblor.

Además, cabe resaltar, que estas correlaciones involucran el eje temporal permitiendo una mejor caracterización de la complejidad del movimiento. A pesar de que estas activaciones pueden tener patrones relevantes que soportan el diagnóstico clínico, pueden estar separados entre el banco de activaciones. Esto puede ser tedioso entonces para comparar durante la interpretación, debido a que se tendrían $|L|$ mapas de activación, para una única probabilidad de salida.
Como alternativa, el presente trabajo propone un *sum-pooling* (resumen de las activaciones por sumas locales entre los filtros). En este caso, la estrategia en mención suma las entradas preservando las dimensiones espaciales en cada capa convolucional. *Sum pooling* reduce la dimensión de los datos mediante la suma de los filtros que contienen la representación de los datos de entrada. La figura 7 ilustra el resultado de la sumatoria de todos los mapas de caracteristicas de una determinada capa convolucional para un ejemplo especifico de una entrada a la CNN propuesta. Particularmente, el valor de activación y_{ij} en el mapa convolucional de suma (MCS), en la posición (i, j) es definido como $y_{ij} = \sum_{k=1}^{C} x_{ij'}^{k}$ donde $x_{ij'}^{k}$ es la activación en la posición (i, j) en el *k-esimo* mapa de características y C es el número de mapas de características en una capa convolucional. El MCS conserva la información espacial porque la operación de agrupación se lleva a cabo en todos los mapas de características, mientras que las otras operaciones de *pooling* tradicionales agregan un mapa de características a cada una de las características.

4.3.2. Grad-CAM: Como segunda estrategía de explicabilidad, en este trabajo se adaptaron los mapas Grad-CAM (*Gradient-weighted Class Activation Mapping*) a las respuestas volumétricas de las diferentes capas convolucionales en 3D, de la arquitectura propuesta. Esta estrategia genera un mapa de calor que destaca las regiones importantes para determinada entrada de la red mediante la información del gradiente que fluye desde una determinada probabilidad de salida hacia una capa convolucional especifica de la CNN propuesta. Midiendo así la importancia de cada canal sobre cada capa en la decisión tomada. Generalmente, se toma la capa convolucional, que está más cercana con el volumen de entrada, para obtener una mejor visualización de la salida ya que las representaciones más profundas de la red capturan mejor la información de alto nivel, perdiendo detalles espaciales.

Figura 7. Estrategia de agrupación de suma para mapas de características en una capa convolucional. a) Imagen de entrada b) Mapas de características de una determinada capa convolucional c) Sumpooling



Para obtener el mapa de localización discriminante de ancho u y alto v asociada a cualquier clase c, y un video de entrada, en primer lugar, se calcula el gradiente del *score* que la red asigna a dicha clase, y^c con respecto a los mapas de características A^k de una capa convolucional. Estos gradientes que fluyen hacia atrás se promedian espacialmente para obtener los pesos α_k^c

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

Donde $\frac{1}{Z} \sum_i \sum_j$ es el *global average pooling* y $\frac{\partial y^c}{\partial A_{ij}^k}$ son los *gradients via backpropagation*. Después de calcular los pesos para la clase objetivo c, se realiza una combinación ponderada de los mapas de activación, seguida de una unidad lineal rectificada (ReLU). Se aplica ReLU a la combinación lineal porque solo interesa las características que tienen una influencia positiva en la clase de interés. Sin ReLU, el mapa de activación de clases resalta más de lo necesario y, por lo tanto, logra un bajo rendimiento de localización.

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k) \tag{2}$$

donde $(\sum_k \alpha^c_k A^k)$ es la combinación lineal. Finalmente, se hace un *upsampling* del resultado garantizando la misma resolución de la imagen original y se muestra en forma de mapa de calor con el fin de ser más intuitivo.

En las Figuras 8 y 9 se evidencia la evolución jerárquica de los mapas de calor a través de las capas que componen la arquitectura convolucional propuesta.

Figura 8. Representación de los mapas de calor generados por el método Grad-CAM en todas las capas convolucionales. La fila superior evidencia los mapas de calor asociados a un paciente con Parkinson, mientras que la fila inferior para un paciente control, ambas para configuraciones posturales.

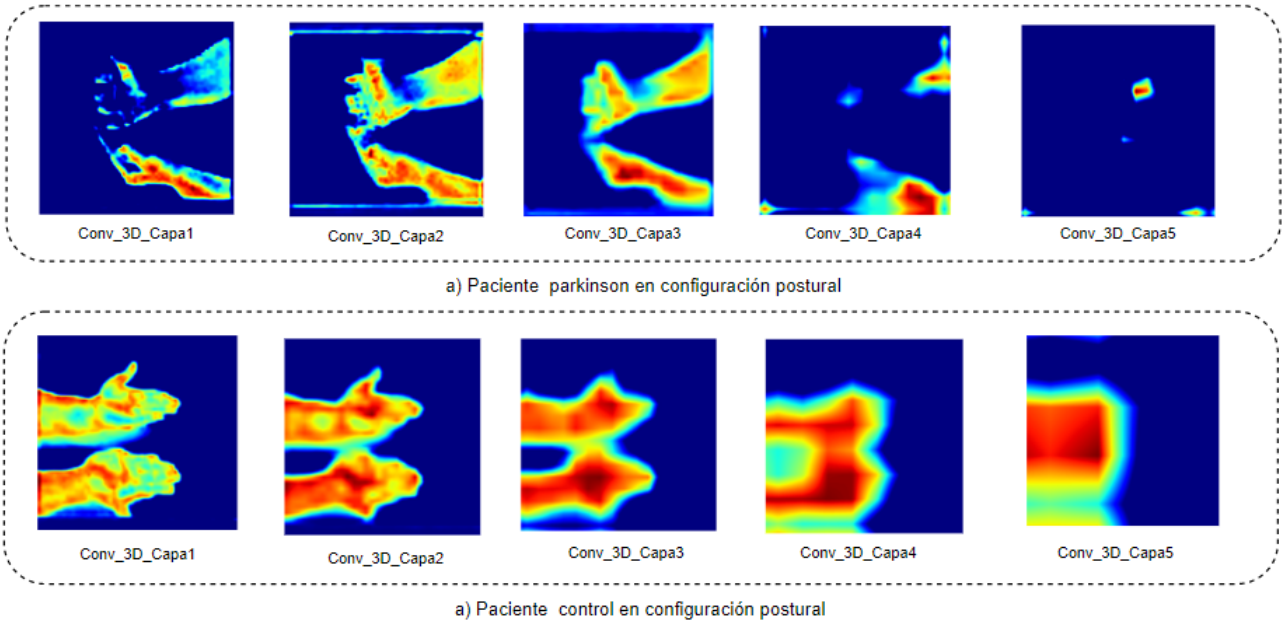
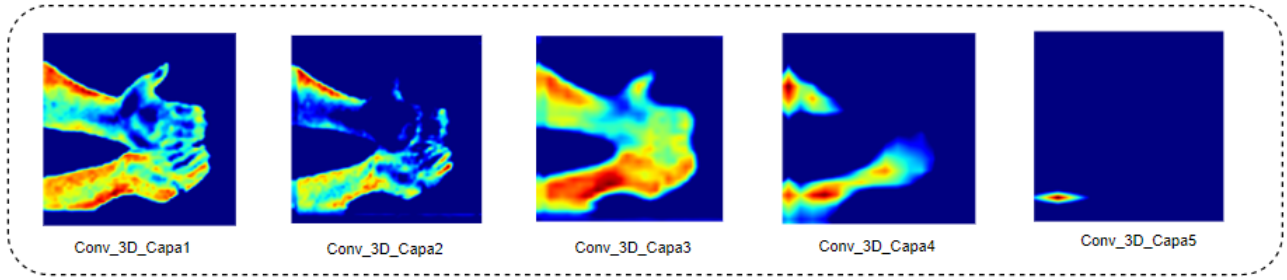
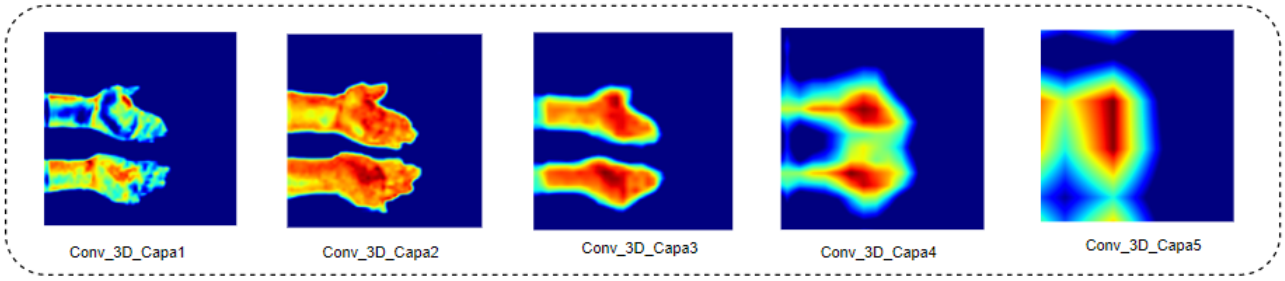


Figura 9. Representación de los mapas de calor generados por el método Grad-CAM en todas las capas convolucionales. La fila superior evidencia los mapas de calor asociados a un paciente con Parkinson, mientras que la fila inferior para un paciente control, ambas para configuraciones de reposo.



a) Paciente parkinson con configuración de reposo



a) Paciente control con configuración de reposo

5. DISEÑO EXPERIMENTAL

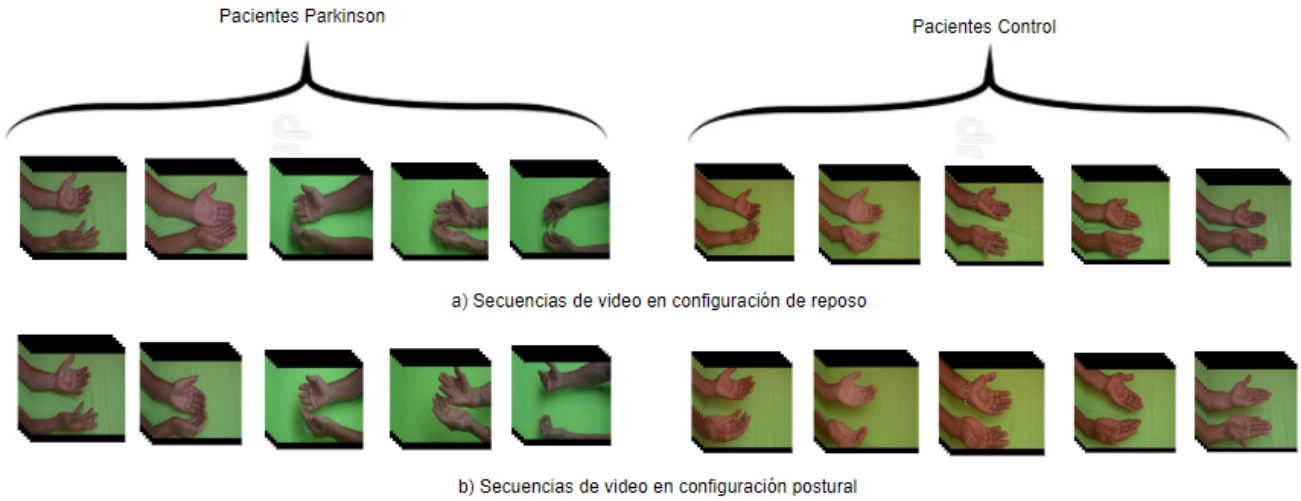
5.1. CONJUNTO DE DATOS

El conjunto de datos con el que se desarrolló la presente propuesta investigativa consiste en videos de 5 pacientes diagnosticados con EP y 5 pacientes control. Los pacientes parkinsonianos fueron diagnosticados previamente en estadios 2 y 3 de la enfermedad por un experto usando protocolos típicos mediante la rutina clínica. Cada paciente se grabó 4 veces tanto para las configuraciones de reposo como postural. En total, el conjunto de datos contiene 40 secuencias de video (estándar) que registran los temblores de las manos en posiciones posturales y de reposo. Se realizó el procesamiento de los datos que consistio en recortar los videos así como fijar los periodos temporales y también se válido con versiones magnificadas de los videos. Para la magnificación de los videos se utilizó una estrategia Euleriana, la cual consiste en amplificar un conjunto de frecuencias temporales de video a través de una proyección multiescala, aplicando una descomposición óptica que permite resaltar las frecuencias de movimiento con mayor relación al fenómeno del temblor obteniéndose un total de 80 videos.

Para cada grabación los sujetos objeto de estudio tenían que mantener las palmas de las manos hacia arriba en la posición más relajada posible durante aproximadamente 12 a 15 segundos. Para la configuración en reposo, las manos debían estar apoyadas y para la configuración postural las manos debian estar levantadas. En cuanto al escenario, se utilizó una cámara puesta estáticamente a 45 grados de un trípode y se empleó un fondo verde para destacar las manos del fondo. Adicionalmente, se tuvo en cuenta un ambiente semicontrolado para evitar artefactos de luminancia externa. En [36] se muestran más detalles del protocolo de captura.

Este estudio fue aprobado por el Comité de Ética de la Universidad Industrial de Santander (UIS) y se obtuvo el consentimiento informado por escrito de todos los pacientes. Los datos

Figura 10. Muestras del conjunto de secuencias de videos de temblor de manos utilizado para el desarrollo de este trabajo. En la gráfica se observa un video por configuración y por paciente.



registrados fueron posibles gracias a la fundación FAMPAS (Fundación del Mayor de Adultos y Parkinson Santander) y al grupo de investigación BIVL2ab (Biomedical Imaging, Vision and Learning Laboratory).

5.2. CONFIGURACIÓN DE LA ESTRATEGIA

5.2.1. Estrategia profunda convolucional 3D: En la tabla 1se resume la arquitectura implementada en este trabajo, para realizar la validación y evaluación de resultados. En general la red contiene 5 convoluciones espacio temporales, el número de los filtros para cada convolución son de 64, 128, 256, 256 y 256 respectivamente. Seguido tiene dos capas totalmente conectadas de dimensión 512 y finalmente una *softmax* que permite obtener la predicción de las etiquetas representado por el número de clases, en este caso, binario.

En total, la configuración propuesta contiene al rededor de 6 millones de parámetros. El tamaño del *kernel* espacio-temporal es de $3 \times 3 \times 3$ con pasos de $1 \times 1 \times 1$ para cada una de las convoluciones, a diferencia de la última, la cual tiene pasos de $1 \times 2 \times 2$. Cada una de estas convoluciones está acompañada por una activación ReLU y un *Max pooling 3D*, donde el tamaño de la primera

convolución es $1 \times 2 \times 2$ y para las demás es de $2 \times 2 \times 2$ teniendo en cuenta pasos del mismo valor, respectivamente. Así mismo, el tamaño de la salida de cada capa convolucional es el mismo para cada una de ellas. Por otro lado, las capas totalmente conectadas tienen una activación ReLU acompañadas y que debido a la gran cantidad de parámetros, en busca de prevenir el sobre aprendizaje, una tasa de 0.5 se establece como *dropout* siendo un mecanismo efectivo para tal fin. La última, como recién se mencionó, contiene el número de clases (Parkinson, Control) con una activación *softmax* que brinda las probabilidades asignadas por la red sobre cada posible entrada.

Tabla 1. Parámetros de la Arquitectura profunda convolucional 3D

Capas	**Salida**
Input	(12, 128, 128, 3)
Conv_3D_Capa1	(12, 128, 128, 64)
Activation	(12, 128, 128, 64)
Maxpooling	(12, 64, 64, 64)
Conv_3D_Capa2	(12, 64, 64, 128)
Activation_1	(12, 64, 64, 128)
Maxpooling_1	(6, 32, 32, 128)
Conv_3D_Capa3	(6, 32, 32, 256)
Activation_2	(6, 32, 32, 256)
Maxpooling_2	(3, 16, 16,256)
Conv_3D_Capa4	(3, 16, 16, 256)
Activation_3	(3, 16, 16, 256)
Maxpooling_3	(1, 8, 8, 256)
Conv_3D_Capa5	(1, 4, 4, 256)
Activation_4	(1, 4, 4, 256)
Maxpooling_4	(1, 2, 2, 256)
Flatten	(1024)
Dense	(512)
Dropout	(512)
Dense_1	(512)
Dropout_1	(512)
Dense_2	(2)

Se seleccionaron 8, 12, 16 y 24 cuadros para todos los estudios. Es decir, para cada uno de los experimentos realizados, la entrada de la red fue variable con respecto a la cantidad de cuadros.

Así mismo, las mejores configuraciones experimentalmente fueron un *batch* de la misma cantidad de datos de entrada, 20 épocas y una taza de aprendizaje (*learning rate*) de 0,0001 permitiendo controlar qué tanto se ajustan los pesos en la red con respecto a la pérdida del gradiente. De igual manera, se utilizó el optimizador de estimación adaptativa de momentos (*Adam*, por sus siglas en inglés), el cual mantiene un factor de entrenamiento por parámetro que se ve afectado por la media del momentum del gradiente.

5.3. VALIDACIÓN ESTADÍSTICA

La estrategia propuesta fue validada bajo la técnica *leave one patient out* (dejar un paciente fuera), donde por cada iteración un paciente servía con propósitos evaluativos mientras que los restantes desempeñaban el rol de entrenamiento. Esta configuración fue elegida debido a la poca cantidad de datos con la que se realizaron los experimentos. En cuanto a las métricas para definir la validación, se utilizaron puntajes relacionados con la clasificación como la exactitud, la precisión , la sensibilidad y la curva ROC, como se describe a continuación.

5.3.1. Curva ROC: La curva ROC (*Receiver Operating Characteristic*) y su área bajo la curva AUC (*Area Under Curve*) son unos de los métodos más importantes de validación para problemas de clasificación. Al trazar la tasa positiva verdadera (sensibilidad) frente a la tasa de falsos positivos (1 - especificidad), se obtiene la curva ROC. Esta curva permite visualizar el equilibrio entre la tasa de verdaderos positivos y la tasa falsos positivos a medida que se varia unos umbrales definidos entre 0 y 1 que pueden ser interpretados como de "falsa alarma".
Se tienen en cuenta los siguientes términos que indican una contabilización de las predicciones para clasificación binaria:

- Verdaderos positivos: Resultados que se predijeron correctamente como positivos. En este trabajo corresponde a pacientes con Parkinson clasificados como pacientes con Parkinson.
- Falsos positivos: Resultados que se predijeron incorrectamente como positivos. En este trabajo corresponde a pacientes control clasificados como pacientes con Parkinson.

- Verdaderos negativos: Resultados que se predijeron correctamente como negativos. En este trabajo corresponde a pacientes control clasificados como pacientes con control.
- Falsos negativos: Resultados que se predijeron incorrectamente como negativos. En este trabajo corresponde a pacientes con Parkinson clasificados como pacientes control.

La tasa de falsos positivos se calcula como el número de positivos verdaderos divididos entre el número de positivos verdaderos y de falsos negativos. Describe que tan bueno es el modelo prediciendo las clases positivas cuando la salida real es positiva. También se conoce esta tasa como sensibilidad. La tasa de falsos positivos se calcula como el número de falsos positivos dividido entre la suma de falsos positivos con los verdaderos negativos. Por otra parte, la especificidad es la inversa de la tasa de falsos positivos. Se obtiene dividiendo el número total de verdaderos negativos entre la suma de los verdaderos negativos y los falsos positivos.
El área bajo la curva (AUC) puede ser utilizado como resumen de la calidad de clasificación del modelo. Según se desplaza la curva hacia la esquina superior izquierda del gráfico, la calidad del modelo va aumentando. Esto se debe a que mejora en su tasa de verdaderos positivos, minimizando también la tasa de falsos positivos. El valor AUC se utiliza como resumen del rendimiento del modelo. Cuanto más esté hacia la izquierda la curva, más área habrá contenida bajo ella y por ende, mejor será el clasificador.

5.3.2. Curva de la precisión y la sensibilidad: La precisión se calcula como el número de verdaderos positivos entre la suma de verdaderos positivos y de falsos positivos. Describe que tan bueno es el modelo a la hora de predecir las salidas de la clase positiva. La sensibilidad (*recall*) se calcula como el número de verdaderos positivos divididos entre la suma de verdaderos positivos y de falsos positivos. Como resumen, la curva de precisión-sensibilidad enfrenta la precisión (eje y) con la sensibilidad (eje x) para diferentes umbrales de corte. Esta métrica evalúa la similitud entre pacientes Parkinson y sujetos control teniendo en cuenta los valores de precisión y sensibilidad (recall).
El área bajo esta curva define el valor de la métrica (AUPRC) midiendo así el balance de la

precisión y la sensibilidad (recall). Una característica de AUPRC es que no utiliza el valor de verdaderos negativos y debido a ello esta métrica no está envuelta en gran proporción por los valores negativos presentes en los datos, centrándose en el manejo de los ejemplos positivos, que en el caso de este trabajo son secuencias de video de pacientes Parkinson clasificados como pacientes con Parkinson. Si el modelo predice correctamente los ejemplos positivos, AUPRC tendrá un valor alto. Por lo contrario, si el modelo predice incorrectamente los ejemplos positivos, el AUPRC tendrá un valor bajo.

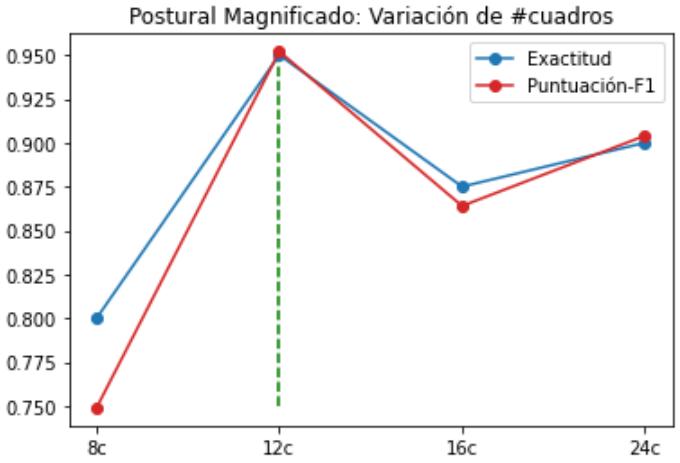
6. EVALUACIÓN Y RESULTADOS

El enfoque propuesto realiza una clasificación y representación de patrones de temblor basada en secuencias de videos capturadas sobre las manos en posición postural y en reposo. En este trabajo, para la arquitectura propuesta se asume que la información más relevante del temblor es encontrada primordialmente de manera temporal. Por lo tanto, esta caracterización temporal permite la representación de patrones de temblor Parkinsoniano basado en la diferenciación del movimiento de manos entre sujetos control y pacientes con Parkinson. Para todos los experimentos resumidos a continuación se consideran secuencias de video en la entrada de la arquitectura. En este caso, se consideraron dos versiones crudas de la secuencia de video y secuencias magnificadas. En cuanto a las secuencias magnificadas, son videos que son primero mapeadas por un filtro temporal para aumentar ciertas frecuencias a través del video, las cuales pueden tener mayor correlación con la enfermedad. En este trabajo se tomó una implementación clásica de Euler Hao-Yu Wu y col. "Eulerian Video Magnification for Revealing Subtle Changes in the World". En: *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)* 31.4 (2012) como etapa de preprocesamiento de los videos, sin variación ni ajuste de ninguno de sus parámetros. A continuación se presentarán los resultados de la estrategia realizada en la metodología propuesta.

Inicialmente se seleccionaron el número de cuadros (frames) considerados por video ubicados equitativamente en el tiempo para determinar el mejor intervalo temporal que capture los patrones motores del temblor en manos. La figura 11 muestra el desempeño del modelo al variar la cantidad de cuadros por video. Específicamente, esta gráfica resume el desempeño de la arquitectura convolucional considerando 8, 12, 16 y 24 frames por video. Se observan resultados similares con excepción de 8 cuadros por video, el cual corresponde a la selección de un frame cada 1.5 segundos para videos con duración de 12 segundos. De esta manera, para intervalos de tiempo entre 1 segundo (12 frames por video) e intervalos de tiempo de 0.5 (24 frames por video) se puede extraer mayor información temporal a partir de la arquitectura.

A continuación todos los resultados son realizados considerando 12 cuadros por video debido a que este fue el mejor resultado obtenido de la figura 11.

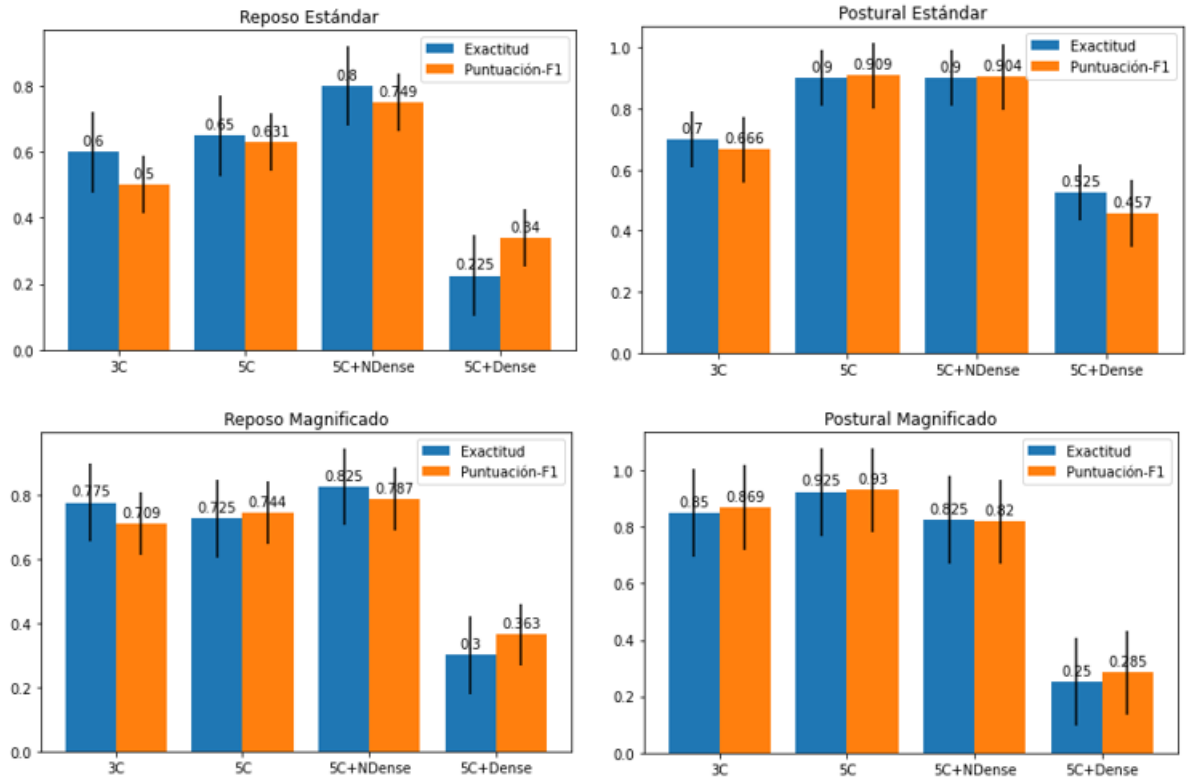
Figura 11. Variación de cuadros para la mejor configuración de la arquitectura. El eje x hace referencia a la variación de los cuadros siendo 8, 12, 16 y 24 respectivamente. El eje y hace referencia a los valores de exactitud (línea azul) y puntuación-F1 (línea roja)



Posteriormente, se analiza la contribución espacio-temporal al considerar 3 y 5 capas convolucionales 3D. En la Figura 12 se muestran las abreviaturas correspondientes a una configuración especifica, 3C (3 capas convolucionales 3D), 5C (5 capas convolucionales 3D), 5C+NDense (5 capas convolucionales 3D y más neuronas a las capas densas, especificamente se hizo la variación cambiando de 512 a 2048 neuronas) y finalmente 5C+Dense (5 capas convolucionales 3D y más capas densas, haciendo referencia a un mayor número de capas densas, de 2 pasar a 4, con 512 neuronas respectivamente). Los mejores resultados de la figura 12 están relacionados con el uso de redes convolucionales más profundas (5 capas) independientemente del movimiento (estándar o magnificado) y de la posición de las manos (reposo o postural). En este sentido la mayor profundidad de la red brinda una mayor información temporal sacrificando información espacial lo que permite relacionar la hipótesis inicial de que la discriminación de los patrones motores está relacionada principalmente con la información temporal respecto de la información espacial.

Adicionalmente, la tabla 2 muestra un detallado análisis que incluye métricas como la exacti-

Figura 12. Representación de resultados de experimentación para diferentes variaciones de capas convolucionales, neuronas y capas densas. A la izquierda gráficas de resultados para configuración en reposo y a la derecha gráficas de resultados para configuración postural.



tud, puntuación-F1, precisión y sensibilidad para los movimientos estándar, magnificados y las configuraciones postural y reposo. De esta manera se observa una complementariedad entre la posición de las manos y el tipo de movimiento (estándar o magnificado). La mayor precisión está relacionada con la posición de reposo sin aumento del movimiento. Esta posición permite un mayor acierto de personas con la enfermedad clasificadas correctamente. Sin embargo, los resultados con mayor Puntuación-F1, sensibilidad y exactitud están relacionados con la posición postural y la magnificación del movimiento. Estos resultados muestran que la información temporal del movimiento de diferentes configuraciones de las manos usadas en el análisis clínico pueden ser cuantificadas adecuadamente mediante el método propuesto.

En términos cuantitativos también se calculó la curva ROC, así como también se calculó la

Tabla 2. Resultados con la mejor arquitectura en diferentes configuraciones

	Exactitud	**Puntuación-F1**	**Precisión**	**Sensibilidad**
Postural Estándar	0,700	0,666	0,750	0,600
Reposo Estándar	0,900	0,888	**1,000**	0,800
Postural Magnificado	**0,925**	**0,930**	0,869	**1,000**
Reposo Magnificado	0,675	0,697	0,652	0,750

precisión y sensibilidad para diferentes umbrales de discriminación. La Figura 13 muestra las curvas ROC y la curvas de precisión-sensibilidad para experimentos en configuración postural con videos magnificados y para experimentos con videos estándar en configuración de reposo, respectivamente. Como se observa, las gráficas c) y d) obtienen un área bajo la curva del 100 % siendo coherente con los resultados evidenciados en la Tabla 2. En todas las ilustraciones se evidencia un alto poder discrimintativo de las predicciones generadas por la arquitectura propuesta. Siendo que para diferentes umbrales se mantienen altas capacidades de discriminación.

En una segunda fase de validación y evaluación de la metodología propuesta, se calcularon mapas que permiten interpretar y soportar las predicciones relacionadas con el Parkinson. En este sentido, la manera más simple de explicar en que se enfoca la red es a través de los mapas de características, considerando que las primeras capas están relacionadas con características de forma, textura, bordes y las capas más profundas presentan un significado más semántico y una visualización más abstracta. En la Figura 14 se evidencia que cuando las entrada de la red son videos estándar independiente de la configuración (reposo o postural) la atención de la red se centra en las zonas de los dedos sin mostrar mayor relevancia (zonas rojas). De manera comparativa en la Figura 15 que hace referencia a entradas de la red con videos magnificados, se puede observar que la atención es mucho más notoria, destacándose por ejemplo que en los mapas de respuesta para la tercera capa convolucional tiene mayor presencia de zonas rojas, lo que quiere decir que el movimiento es mucho más evidente. Este resultado resulta predominante para determinar en una implementación clínica, se debe tener en cuenta videos magnificados

como entradas que dan mayor aporte a la explicabilidad e interpretabilidad de los modelos. Ambas configuraciones (postural y reposo) presentan mayor atención en los bordes de los dedos. Análogamente para los pacientes con Parkinson la mayor atención es prestada en los bordes del dedo pulgar (regiones rojas donde en los videos se presenta el temblor). Sin embargo, en esta representación no se considera la complementariedad entre mapas de características. Representaciones más robustas que consideran varios mapas de características corresponde a los mapas de explicabilidad cuyos resultados de visualización se analizan a continuación.

Estas observaciones son aún más claras desde los mapas de agrupación usando sumas entre filtros (Sum-pooling). Para este enfoque se visualizaron los mapas de agrupación de sumas para la tercera y cuarta capa convolucional de la estrategia propuesta con el modelo que obtuvo los mejores resultados de clasificación. En la Figura 16 se observa que la mayor atención prestada por los mapas de los sujetos control se centra en las palmas de las manos mientras que en los pacientes Parkinson se centra en los bordes de las manos. Por otro lado, usamos un segundo método más robusto y mas claro para la interpretación de la representación.

Así pues, los mapas de representación basados en Grad-CAM son los que presentan la visualización con mayor interpretabilidad donde los colores más cálidos representan una mayor atención (ver Figura 17). Interesantemente, los sujetos control prestan su atención principalmente en toda la mano mientras que los sujetos Parkinson enfocan su atención principalmente en regiones de la mano en las cuales se percibe un temblor sutil poco perceptible durante los videos. En este sentido el método nos muestra como la red identifica las regiones más sensibles al movimiento del paciente (punta de los dedos y región de agarre del dedo pulgar) tanto en la configuración de reposo y postural.

Figura 13. A la izquierda gráficas de la curva ROC en donde el eje x representa la tasa de falsos positivos y el eje y la tasa de verdaderos positivos. A la derecha gráficas de la curva de precisión-sensibilidad en donde el eje x representa la sensibilidad y el eje y la precisión. La gráficas a) y b) corresponden a los resultados del experimento que obtuvo mejores resultados de Exactitud, Puntuación F1 y sensibilidad, siendo Postural Magnificado haciendo referencia a videos magnificados en secuencias de video de temblor de manos en configuración de postural. Las gráficas c) y d) corresponden al mejor resultado del experimento que obtuvo mejores resultados de Precisión, siendo Estándar Reposo haciendo referencia a videos estándar en configuración de reposo.

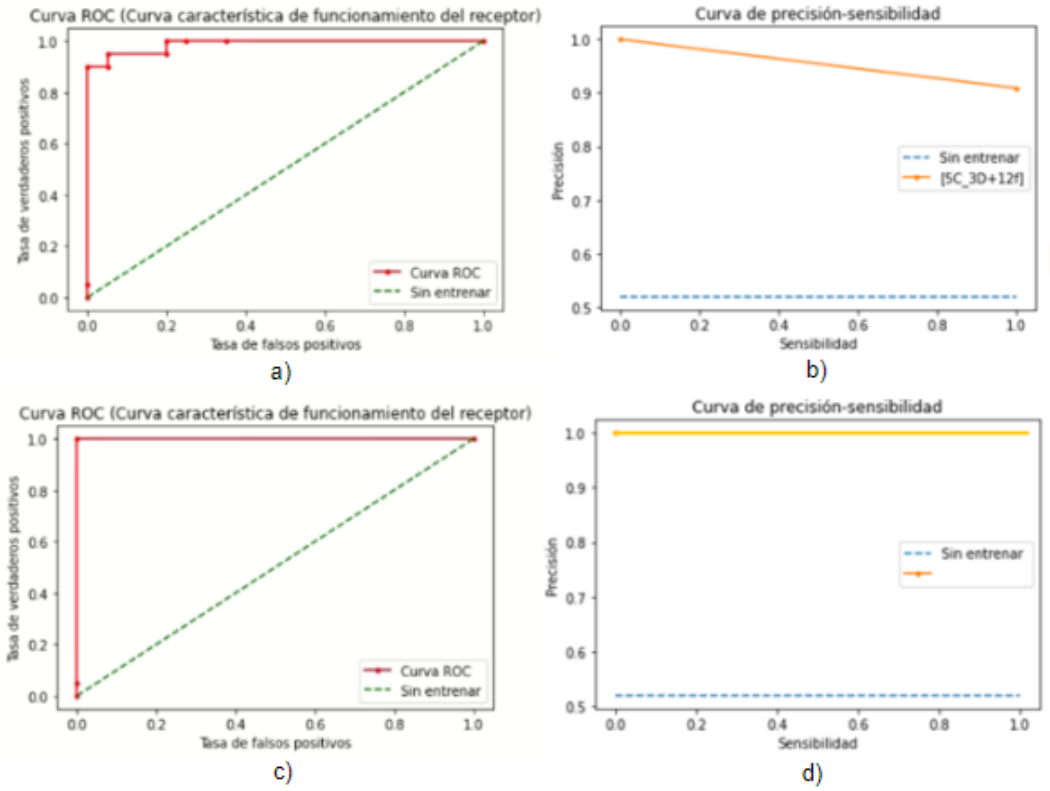
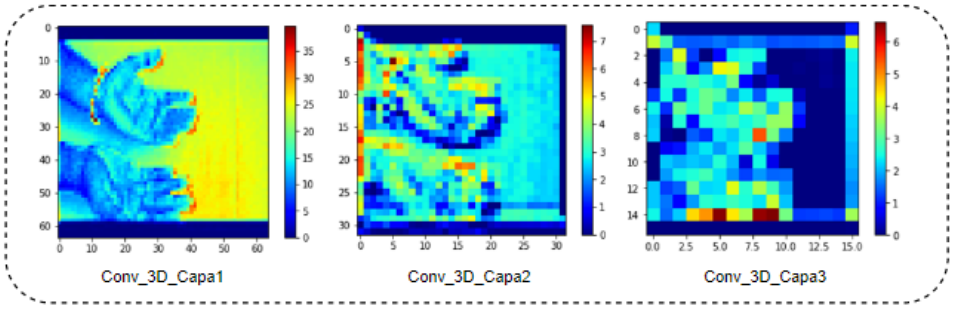
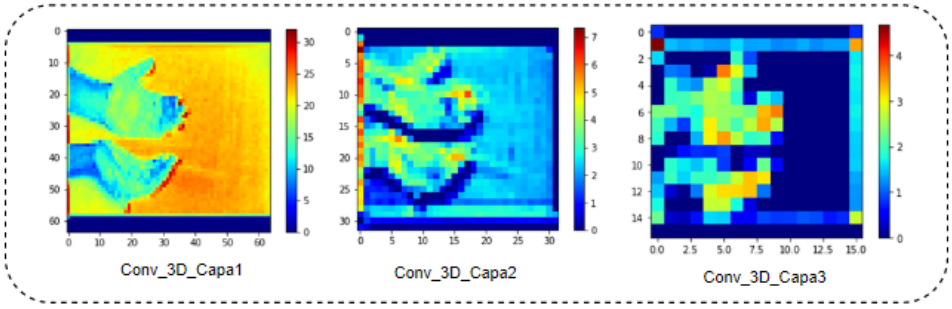


Figura 14. Comparación entre mapas de características en secuencias de videos estándar.

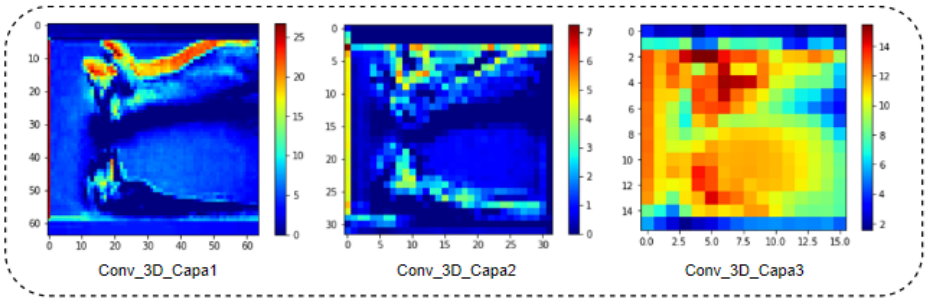


a) Secuencias de videos estándar de un paciente control en configuración postural

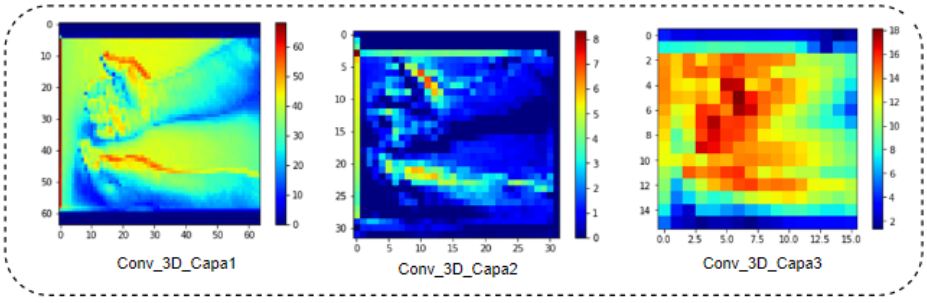


a) Secuencias de videos estándar de un paciente control en configuración de reposo

Figura 15. Comparación entre mapas de características en secuencias de videos magnificados.



a) Secuencias de videos magnificados de un paciente con EP en configuración postural



a) Secuencias de videos magnificados de un paciente con EP en configuración de reposo

Figura 16. Mapas de respuesta con el método sum-pooling propuesto. A la izquierda como se evidencia en la figura, pacientes Parkinson y a la derecha pacientes Control. a) Entrada b) Mapas de respuesta tercera capa convolucional c) Mapas de respuesta para cuarta capa convolucional

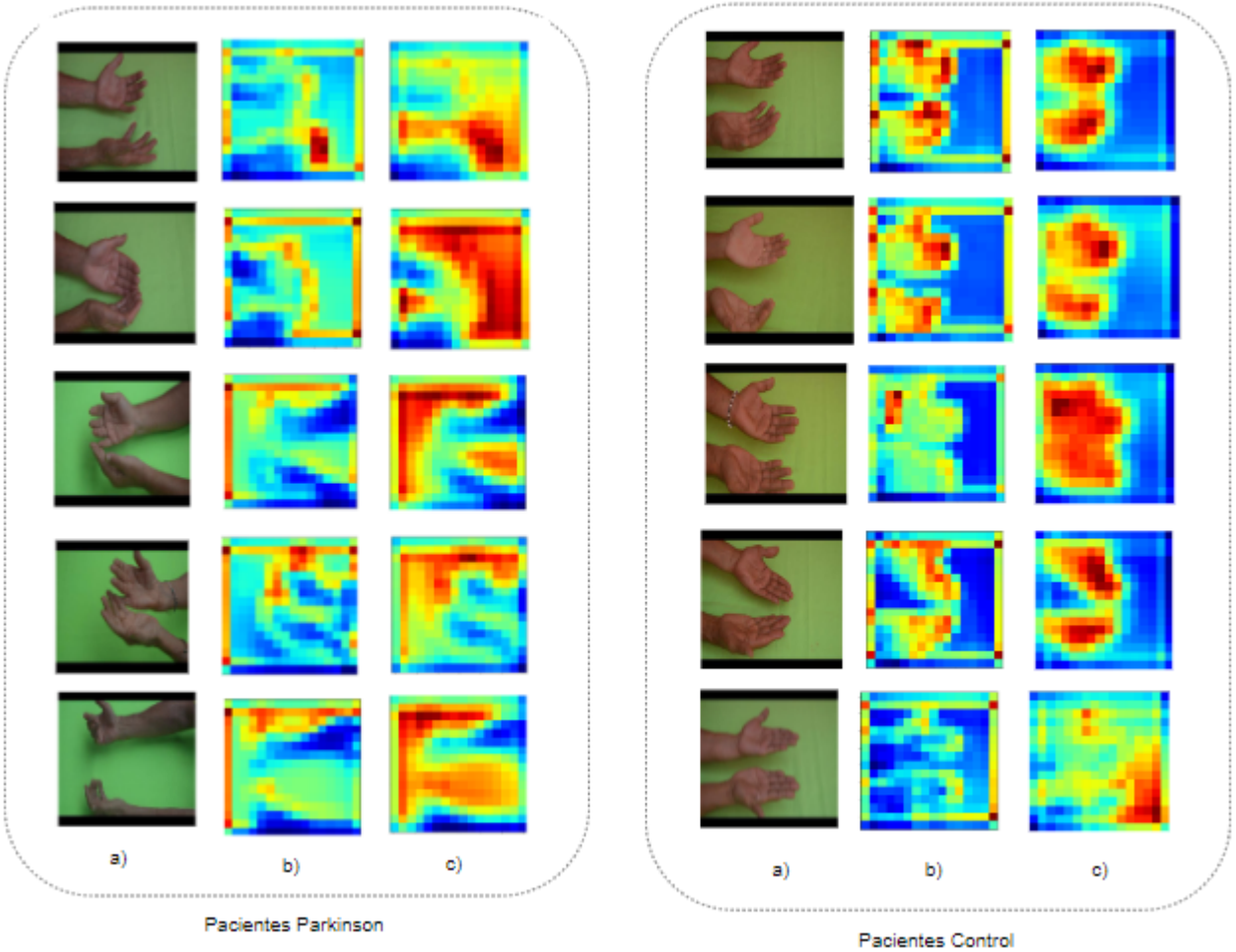
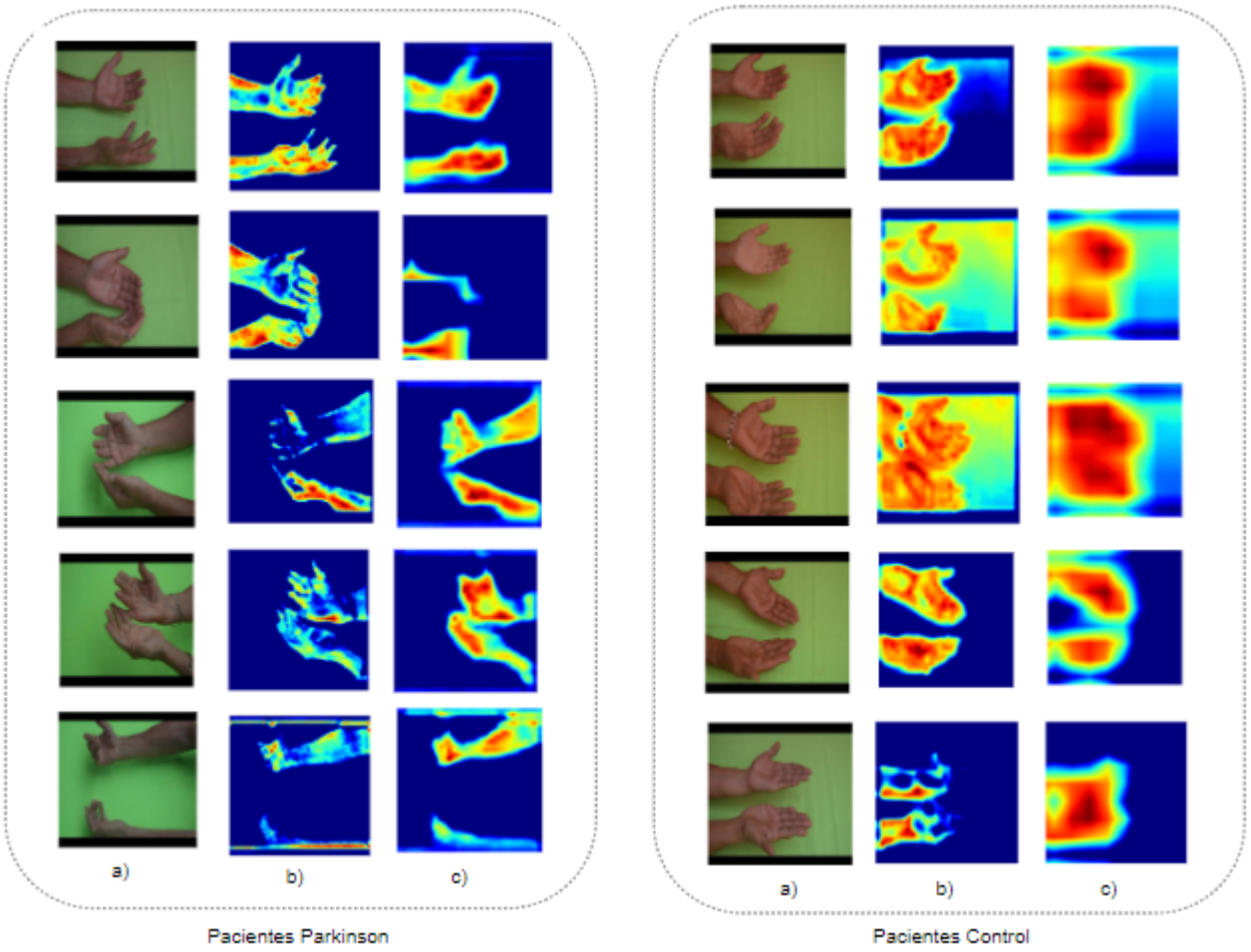


Figura 17. Mapas de calor con el método Grad-CAM propuesto. A la izquierda como se evidencia en la figura, pacientes Parkinson y a la derecha pacientes Control. a) Entrada b) Mapas de respuesta tercera capa convolucional c) Mapas de calor para cuarta capa convolucional



7. CONCLUSIONES Y TRABAJO FUTURO

En el presente trabajo se desarrolló una estrategia convolucional 3D que permite la clasificación y representación visual del temblor de manos asociados a pacientes con Parkinson. En primera medida, se procesó un conjunto de videos relacionados con el temblor de manos en posiciones usuales de la rutina clínica como lo son la posición en reposo y la posición postural. El método propuesto permitió la discriminación entre patrones motores asociados a la Enfermedad de Parkinson y entre diferentes posturas usadas en la rutina clínica permitiendo relacionar una mayor discriminación temporal respecto a la espacial a medida que aumenta la profundidad de la red.

El enfoque mostró resultados relevantes en cuanto a la localización de los patrones de temblor ya que permitió cuantificar información de movimiento preservando la representación espacial y temporal. La estrategia permite visualizar patrones de movimiento a partir de secuencias de video en donde se pudo evidenciar una diferenciación visual entre los mapas de explicabilidad de sujetos control y pacientes con Parkinson. En cuanto al análisis de las configuraciones tanto en reposo como postural del temblor de manos resultaron ser complementarias porque ambas brindan información discriminativa acerca de los patrones motores del temblor en manos.

El presente trabajo tiene un amplio potencial para ser implementado como herramienta de soporte en diferentes etapas de la enfermedad de Parkinson en donde el temblor de manos es apenas perceptible. Sin embargo, el enfoque requiere una amplificación del número de pacientes y el número de videos. Adicionalmente, requiere un procesamiento extra de los videos relacionados con su magnificación. En futuros trabajos la ampliación del conjunto de datos (dataset) y la fusión de estos dos tipos de posiciones (postural y reposo) puede mejorar la aplicabilidad del método en escenarios clínicos.

BIBLIOGRAFÍA

Abboud, Hesham, Anwar Ahmed y Hubert H Fernandez. "Essential tremor: choosing the right management plan for your patient". En: *Cleve Clin J Med* 78.12 (2011), págs. 821-8 (vid. pág. 24).

Anouti, Ahmad y William C Koller. "Tremor disorders. Diagnosis and management." En: *Western journal of medicine* 162.6 (1995), pág. 510 (vid. pág. 23).

Bacher, M, E Scholz y HC Diener. "24 hour continuous tremor quantification based on EMG recording". En: *Electroencephalography and clinical neurophysiology* 72.2 (1989), págs. 176-183 (vid. pág. 27).

Castro Toro, Aracelly y Omar Freddy Buriticá. "Parkinson's disease: diagnostic criteria, risk factors and progression, and assessment scales clinical stage". En: *Acta Neurológica Colombiana* 30.4 (2014), págs. 300-306 (vid. pág. 12).

Ciresan, Dan Claudiu y col. "Flexible, high performance convolutional neural networks for image classification". En: *Twenty-second international joint conference on artificial intelligence*. 2011 (vid. pág. 16).

Contreras, Sergio, Isail Salazar y Fabio Martínez. "Parkinsonian hand tremor characterization from magnified video sequences". En: *14th International Symposium on Medical Information Processing and Analysis*. Vol. 10975. International Society for Optics y Photonics. 2018, pág. 1097503 (vid. págs. 28, 43).

Doshi-Velez, Finale y Been Kim. "Towards a rigorous science of interpretable machine learning". En: *arXiv preprint arXiv:1702.08608* (2017) (vid. pág. 19).

Elble, Rodger J. “Tremor”. En: *Neuro-geriatrics*. Springer, 2017, págs. 311-326 (vid. pág. 12).

Graves, Alex. “Supervised sequence labelling”. En: *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, págs. 5-13 (vid. pág. 18).

Greffard, Sandrine y col. “Motor score of the Unified Parkinson Disease Rating Scale as a good predictor of Lewy body–associated neuronal loss in the substantia nigra”. En: *Archives of neurology* 63.4 (2006), págs. 584-588 (vid. pág. 26).

Hochreiter, Sepp. “Ja1 4 rgen schmidhuber (1997).“long short-term memory””. En: *Neural Computation* 9.8 () (vid. pág. 18).

Huo, Weiguang y col. “A heterogeneous sensing suite for multisymptom quantification of Parkinson’s disease”. En: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28.6 (2020), págs. 1397-1406 (vid. pág. 25).

Jeyakumar, Jeya Vikranth y col. “How can i explain this to you? an empirical study of deep neural network explanation methods”. En: *Advances in Neural Information Processing Systems* 33 (2020), págs. 4211-4222 (vid. pág. 19).

Kim, Han Byul y col. “Wrist sensor-based tremor severity quantification in Parkinson’s disease using convolutional neural network”. En: *Computers in biology and medicine* 95 (2018), págs. 140-146 (vid. pág. 29).

Koller, William C, Bridget Vetere-Overfield y Ruth Barter. “Tremors in early Parkinson’s disease.” En: *Clinical neuropharmacology* 12.4 (1989), págs. 293-297 (vid. pág. 23).

LeCun, Yann, Yoshua Bengio y Geoffrey Hinton. “Deep learning”. En: *nature* 521.7553 (2015), págs. 436-444 (vid. págs. 15, 28).

LeMoyne, Robert, Cristian Coroian y Timothy Mastroianni. "Quantification of Parkinson's disease characteristics using wireless accelerometers". En: *2009 ICME International Conference on Complex Medical Engineering*. IEEE. 2009, págs. 1-5 (vid. pág. 27).

Mader, Malenka y col. "Spectral and higher-order-spectral analysis of tremor time series". En: *Clin Exp Pharmacol* 4.149 (2014), págs. 2161-1459 (vid. pág. 26).

Mansur, Paulo Henrique G y col. "A review on techniques for tremor recording and quantification". En: *Critical Reviews™ in Biomedical Engineering* 35.5 (2007) (vid. págs. 13, 22).

Oliveira Andrade, Adriano de y col. "Task-Specific Tremor Quantification in a Clinical Setting for Parkinson's Disease". En: *Journal of Medical and Biological Engineering* 40.6 (2020), págs. 821-850 (vid. pág. 25).

Pasquini, Jacopo y col. "Progression of tremor in early stages of Parkinson's disease: a clinical and neuroimaging study". En: *Brain* 141.3 (2018), págs. 811-821 (vid. pág. 23).

Patel, Shyamal y col. "Using wearable sensors to predict the severity of symptoms and motor complications in late stage Parkinson's Disease". En: *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2008, págs. 3686-3689 (vid. pág. 27).

Poewe, Werner y col. "Parkinson disease". En: *Nature reviews Disease primers* 3.1 (2017), págs. 1-21 (vid. pág. 22).

Riviere, Cameron N, Stephen G Reich y Nitish V Thakor. "Adaptive Fourier modeling for quantification of tremor". En: *Journal of neuroscience methods* 74.1 (1997), págs. 77-87 (vid. pág. 26).

Rodriguez-Blazquez, Carmen, Maria João Forjaz y Pablo Martinez-Martin. "Rating scales in movement disorders". En: *Movement Disorders Curricula*. Springer, 2017, págs. 65-75 (vid. pág. 26).

Selvaraju, Ramprasaath R y col. "Grad-cam: Visual explanations from deep networks via gradient-based localization". En: *Proceedings of the IEEE international conference on computer vision*. 2017, págs. 618-626 (vid. pág. 22).

Sharma, Sushil y col. "Biomarkers in Parkinson's disease (recent update)". En: *Neurochemistry international* 63.3 (2013), págs. 201-229 (vid. pág. 12).

Soran, Bilge y col. "Tremor detection using motion filtering and SVM". En: *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE. 2012, págs. 178-181 (vid. pág. 28).

Timmer, Jens y col. "Characteristics of hand tremor time series". En: *Biological cybernetics* 70.1 (1993), págs. 75-80 (vid. pág. 26).

Uhríková, Zdenka y col. "Action tremor analysis from ordinary video sequence". En: *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2009, págs. 6123-6126 (vid. pág. 27).

Vaillancourt, David E y Karl M Newell. "The dynamics of resting and postural tremor in Parkinson's disease". En: *Clinical Neurophysiology* 111.11 (2000), págs. 2046-2056 (vid. pág. 23).

Varol, Gül, Ivan Laptev y Cordelia Schmid. "Long-Term Temporal Convolutions for Action Recognition". En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6 (2018), págs. 1510-1517. DOI: 10.1109/TPAMI.2017.2712608 (vid. pág. 17).

Vickers, Neil J. "Animal communication: when i'm calling you, will you answer too?" En: *Current biology* 27.14 (2017), R713-R715 (vid. pág. 17).

Wirdefeldt, Karin y col. "Epidemiology and etiology of Parkinson's disease: a review of the evidence". En: *European journal of epidemiology* 26.1 (2011), pág. 1 (vid. pág. 12).

Wu, Hao-Yu y col. "Eulerian Video Magnification for Revealing Subtle Changes in the World". En: *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)* 31.4 (2012) (vid. pág. 49).

Xu, Zhenqi, Jiani Hu y Weihong Deng. "Recurrent convolutional neural network for video classification". En: *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2016, págs. 1-6 (vid. págs. 17, 18).

Zhang, Jie y col. "Differential diagnosis of Parkinson disease, essential tremor, and enhanced physiological tremor with the tremor analysis of EMG". En: *Parkinson's Disease* 2017 (2017) (vid. pág. 24).

Zheng, Xiaochen y col. "Activity-aware essential tremor evaluation using deep learning method based on acceleration data". En: *Parkinsonism & related disorders* 58 (2019), págs. 17-22 (vid. pág. 29).

Zhou, Bolei y col. "Learning deep features for discriminative localization". En: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, págs. 2921-2929 (vid. pág. 21).