UNIVERSIDAD DE VALLADOLID

FACULTAD DE MEDICINA

ESCUELA DE INGENIERÍAS INDUSTRIALES

TRABAJO DE FIN DE GRADO

GRADO EN INGENIERÍA BIOMÉDICA

APLICACIÓN DE TÉCNICAS DE *DEEP LEARNING* PARA CLASIFICAR LOS EVENTOS DE APNEA E HIPOPNEA MEDIANTE LAS SEÑALES DE PULSIOXIMETRÍA

Autor:

D.ª Beatriz Pascual Roa

Tutores:

Dr. D. Fernando Vaquerizo Villar

Dr. D. Roberto Hornero Sánchez

Valladolid, 27 de septiembre de 2023

TÍTULO:	**Aplicación de técnicas de *deep learning* para clasificar los eventos de apnea e hipopnea mediante las señales de pulsioximetría.**
AUTOR:	**D.ª Beatriz Pascual Roa**
TUTOR/A:	**Dr. D. Fernando Vaquerizo Villar** **Dr. D. Roberto Hornero Sánchez**
DEPARTAMENTO:	**Teoría de la Señal y Comunicaciones e Ingeniería Telemática**

TRIBUNAL

PRESIDENTE:	**Dr. D. Roberto Hornero Sánchez**
SECRETARIO:	**Dr. D. Daniel Álvarez González**
VOCAL:	**Dr. D.ª María García Gadañón**
SUPLENTE 1:	**Dr. D. Carlos Gómez Peña**
SUPLENTE 2:	**Dr. D. Jesús Poza Crespo**

AGRADECIMIENTOS

Me gustaría comenzar agradeciendo a mis tutores Fernando Vaquerizo Villar y Roberto Hornero Sánchez por su orientación y apoyo incondicional durante la realización de este Trabajo de Fin de Grado. Gracias por brindarme la oportunidad de adentrarme en el ámbito de la investigación y resolver todas las dudas que me han ido surgiendo a lo largo de esta etapa.

También quiero agradecer a todos los miembros del Grupo de Ingeniería Biomédica de la Universidad de Valladolid, y en especial, a mis compañeros de laboratorio. Desde el primer día me acogieron de la mejor manera posible, compartiendo sus conocimientos conmigo y haciendo que estos meses hayan sido mucho más divertidos.

Por último, a mi familia y amigos, les agradezco por todo su cariño y paciencia a lo largo de este tiempo. Especialmente a mis padres, por ser mi mayor apoyo y mi referente siempre.

Muchas gracias a todos.

RESUMEN

La apnea obstructiva del sueño (AOS) es una patología de gran prevalencia en la población general con graves repercusiones para la calidad de vida de las personas que la padecen. Está directamente relacionada con el desarrollo de enfermedades cardiovasculares, además de aumentar el riesgo de accidentes de tráfico y la tasa de mortalidad. A pesar de que la polisomnografía nocturna es reconocida como el *gold standard* para el diagnóstico de la AOS, presenta una serie de limitaciones significativas. Se trata de una prueba con un elevado coste económico, laboriosa y no siempre accesible, aparte de ser incómoda para los pacientes al tener que dormir una noche fuera de sus domicilios particulares conectados a múltiples sensores.

Ante estos inconvenientes, la comunidad científica ha explorado diversas alternativas para ayudar en el diagnóstico de la AOS. Entre ellas se encuentra la pulsioximetría, una técnica simple, fiable y accesible que registra las señales de saturación de oxígeno (SpO_2) y frecuencia de pulso (PR), las cuales contienen información acerca de los episodios de hipoxemia intermitente, normalmente asociados con la aparición de eventos de apnea e hipopnea.

En este contexto, el objetivo principal de este Trabajo Fin de Grado (TFG) ha sido evaluar la utilidad de técnicas de *deep learning* aplicadas sobre las señales de SpO_2 y PR procedentes de la pulsioximetría para la detección automática de eventos de apnea e hipopnea, así como para la estimación del AHI. Para lograr dicho objetivo, se ha utilizado la base de datos MESA, que contiene 2056 registros de SpO_2 y PR de sujetos adultos. Se han entrenado arquitecturas de *deep learning* basadas en redes neuronales convolucionales (CNN) con las señales de SpO_2 y PR por separado, así como con ambas señales combinadas. Además, se exploran diferentes tamaños de segmentos de entrada (30 y 60 segundos) y se prueba una estrategia de adyacencia. Asimismo, se aplica el método *Gradient-weighted Class Activation Mapping* (Grad-CAM), una técnica de *Explainable Artificial Intelligence*, para comprender mejor las decisiones de los algoritmos e identificar los patrones de las señales de SpO_2 y PR relacionados con apneas e hipopneas.

Se ha obtenido un gran rendimiento en la detección de eventos respiratorios, con una exactitud del 85.13% para segmentos de 30 segundos con adyacencia y una exactitud del 83.53% para segmentos de 60 segundos con adyacencia, resaltando la importancia de la información temporal previa. En cuanto a la estimación del AHI, se lograron kappas de 0.582 y 0.576 para segmentos de 30 y 60 segundos, respectivamente, utilizando únicamente la señal de SpO_2. Sin embargo, la señal de PR no ha demostrado ser útil por si sola para la detección de eventos de apnea e hipopnea, ni ha contribuido a mejorar los resultados al ser combinada con la señal de SpO_2 (SpO_2+PR). Por último, los mapas de calor obtenidos mediante Grad-CAM han proporcionado una justificación de las predicciones de la CNN, facilitando la identificación de las características de las señales de SpO_2 y PR relacionadas con los eventos respiratorios. Concretamente, se ha observado que las regiones de interés del modelo se centran en áreas donde ocurren desaturaciones de oxígeno (SpO_2) y variaciones en la frecuencia cardiaca (PR), prestando especial

atención a los puntos mínimos o *nadires* de la señal de SpO_2. Estos resultados obtenidos reflejan la utilidad de los modelos de *deep learning* aplicados a la pulsioximetría en la detección de eventos de apnea e hipopnea, lo cual podría tener un impacto significativo en el diagnóstico de esta patología.

Palabras clave: Pulsioximetría, Detección de eventos, Apnea obstructiva del sueño (AOS), Redes neuronales convolucionales (CNN), *Deep learning* (DL), *Explainable Artificial Intelligence* (XAI).

ABSTRACT

Obstructive sleep apnea (OSA) is a highly prevalent condition in the general population, with significant implications for the quality of life of affected individuals. It is directly associated with the development of cardiovascular diseases, heightened risks of traffic accidents, and increased mortality rates. Despite the widespread recognition of nocturnal polysomnography as the gold standard for diagnosing OSA, it presents several substantial limitations. It is an expensive, labor-intensive, and not always readily accessible test, and proves uncomfortable for patients who must spend a night away from their homes connected to multiple sensors.

In response to these limitations, the scientific community has explored various alternatives for diagnosing OSA. Among them, pulse oximetry is a simple, reliable, and accessible technique that records oxygen saturation (SpO_2) and pulse rate (PR) signals. These signals contain information about episodes of intermittent hypoxemia, which are typically associated with the occurrence of apnea and hypopnea events.

In this context, the main objective of this work has been to evaluate the utility of deep learning techniques applied to the SpO_2 and PR signals obtained from pulse oximetry for the automatic detection of apnea and hypopnea events and, subsequently, the estimation of the apnea-hypopnea index (AHI). To achieve this goal, the MESA database, comprising 2056 recordings of SpO_2 and PR signals from adult subjects, was utilized. Deep learning architectures based on convolutional neural networks (CNNs) were trained with SpO_2 and PR signals individually, as well as with both signals combined. Additionally, various input segment sizes (30 and 60 seconds) were explored, along with an adjacency strategy. Gradient-weighted Class Activation Mapping (Grad-CAM) method, and Explainable Artificial Intelligence technique, was then applied to gain deeper insights into the algorithm's decision-making process and to identify patterns in the SpO_2 and PR signals related to apnea and hypopnea events.

Remarkable performance was achieved in the detection of respiratory events, with an accuracy of 85.13% for 30-second segments with adjacency and 83.53% for 60-second segments with adjacency, underscoring the importance of prior temporal information in enhancing the precision of event detection. Regarding AHI estimation, Cohen's kappa values of 0.582 and 0.576 were achieved for 30 and 60-second segments, respectively, using only the SpO_2 signal. However, the PR signal alone did not prove to be useful for the detection of apnea and hypopnea events, nor did its inclusion alongside the SpO_2 signal (SpO_2+PR) yield improved results. Finally, the heatmaps obtained through Grad-CAM provided justification for the CNN’s predictions, thereby facilitating the identification of SpO_2 and PR signal characteristics related to respiratory events. Notably, Grad-CAM heatmaps revealed that the model’s regions of interest primarily focused on areas marked by oxygen desaturations (SpO_2) and heart rate variations (PR). Particular emphasis was placed on the nadirs of the SpO_2 signal. These results underscore the utility of deep learning algorithms applied to pulse oximetry signals in detecting apnea and hypopnea events, which could have a significant impact on the diagnosis of this condition.

Keywords: Pulse oximetry, Event detection, Obstructive Sleep Apnea (OSA), Convolutional Neural Networks (CNN), Deep learning (DL), Explainable Artificial Intelligence (XAI).

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

CAPÍTULO 1: INTRODUCCIÓN

CAPÍTULO 3: SUJETOS Y SEÑALES

CAPÍTULO 4: METODOLOGÍA

CAPÍTULO 5: RESULTADOS

ÍNDICE DE TABLAS

CAPÍTULO 2: DETECCIÓN DE EVENTOS DE APNEA E HIPOPNEA

CAPÍTULO 3: SUJETOS Y SEÑALES

CAPÍTULO 5: RESULTADOS

CAPÍTULO 6: DISCUSIÓN

CAPÍTULO 1: INTRODUCCIÓN

1.1 APNEA OBSTRUCTIVA DEL SUEÑO

Los trastornos respiratorios durante el sueño presentan una amplia variedad de manifestaciones y, aunque su divulgación es reciente, su existencia no lo es. Entre estos trastornos, destaca la apnea obstructiva del sueño, la cual ya fue documentada en la antigua Grecia. En un texto datado en el año 330 a.C. se describía el caso de un rey de Pontus, que presentaba síntomas característicos de glotonería, obesidad y dificultad para mantenerse despierto, hasta el punto de requerir el empleo de agujas para despertarlo (González Mangado et al., 2020).

A pesar de estos registros históricos, fue en la segunda mitad del siglo XX cuando se logró establecer una definición precisa de esta enfermedad que ha afectado a la humanidad desde hace mucho tiempo. En 1973, Christian Guilleminault acuñó el término "síndrome de apnea del sueño", marcando un hito en la compresión y clasificación de esta condición (Guilleminault et al., 1973).

La apnea obstructiva del sueño (AOS) o por su denominación en inglés, *Obstructive Sleep Apnea* (OSA), se caracteriza por la aparición de una obstrucción del flujo aéreo durante el sueño debido al colapso parcial o completo de las vías respiratorias superiores (VAS), consecuente a un fallo anatomo-funcional de las mismas (Ho & Brass, 2011). Todo ello resulta en la aparición de eventos de apneas, hipopneas, ronquidos, y/o microdespertares (*arousals*) (Cuesta, 2005).

En consecuencia, estos episodios repetidos pueden ocasionar desaturaciones nocturnas, lo que a su vez puede provocar un sueño no reparador, excesiva somnolencia diurna y alteraciones en el sistema cardiovascular, respiratorio y neurocognitivo (Cuesta, 2005).

1.1.1. EPIDEMIOLOGÍA

La AOS es una patología de gran prevalencia en la población general, que conlleva múltiples implicaciones negativas para la calidad de vida de quienes la padecen. Esta condición se asocia directamente con el desarrollo de hipertensión arterial, enfermedades cardiovasculares y cerebrovasculares, así como un incremento en el riesgo de accidentes de tráfico y un aumento en la tasa de mortalidad (Lloberes et al., 2011).

Dado el impacto significativo de esta enfermedad en la sociedad, se considera a la AOS un problema de salud pública de gran magnitud. En España, se estima que entre el 3% y el 6% de la población tiene síntomas de AOS y entre el 24% y el 26% sufren más de 5 eventos de apnea o hipopnea por hora de sueño (Cuesta, 2005; Lloberes et al., 2011). Por otro lado, en EEUU aproximadamente uno de cada cinco adultos presenta al menos AOS leve, mientras que uno de cada quince adultos padece AOS moderado (Ho & Brass, 2011). Debido a la alta prevalencia de la enfermedad, es necesario abordar este problema de manera efectiva y proporcionar un diagnóstico temprano (Ho & Brass, 2011). No

obstante, debido a la falta de concienciación tanto en la población general como en los profesionales de la salud, la gran mayoría de los casos no son diagnosticados ni tratados adecuadamente (Ho & Brass, 2011). Asimismo, los pacientes que no cuentan con un diagnóstico de AOS consumen recursos sanitarios de forma considerablemente mayor en comparación con aquellos que han sido correctamente diagnosticados y reciben tratamiento adecuado (Lloberes et al., 2011). Por tanto, es de gran importancia poder identificar y tratar a los pacientes afectados.

En lo que respecta a los factores de riesgo, la edad, el sexo masculino y, sobre todo, el índice de masa corporal (IMC) son los más importantes (Lloberes et al., 2011). En relación con el IMC, se estima que aproximadamente el 70% de las personas afectadas son obesas, y que la prevalencia de esta condición en hombres y mujeres con obesidad ronda el 40%. Asimismo, se ha observado que un 26% de los pacientes con un IMC superior a 30 y un 33% con IMC superior 40 presentan AOS moderado (Ho & Brass, 2011). Por otra parte, la prevalencia de esta patología se triplica en individuos mayores de 65 años en comparación con aquellos que se encuentran en el rango de edad entre 40 y 65 años (Lloberes et al., 2011).

Con relación al género, los estudios estiman una mayor prevalencia de la AOS en hombres, con una proporción hombre-mujer de aproximadamente 3/1 (Lloberes et al., 2011). Sin embargo, esta proporción se iguala a partir de la menopausia, aumentándose el riesgo de AOS en las mujeres postmenopáusicas en comparación con las premenopáusicas (Ho & Brass, 2011).

Por último, los factores estructurales relacionados con la anatomía ósea craneofacial desempeñan un papel significativo en la predisposición de los pacientes al colapso de la faringe durante el sueño. En los estudios de imagen se evidencia que los pacientes que sufren AOS presentan una disminución en el diámetro faríngeo (Ho & Brass, 2011).

1.1.2. FISIOPATOLOGÍA

La AOS se caracteriza por la presencia de dos procesos fisiológicos principales que desencadenan la sintomatología clínica (Lloberes et al., 2011). En primer lugar, las apneas, hipopneas e hipoxia intermitentes juegan un papel crucial en la manifestación de los síntomas. La *American Academy of Sleep Medice* (AASM) utiliza las siguientes definiciones de apnea e hipopnea (Berry et al., 2022):

- **Apnea**: reducción del flujo aéreo (FA) del 90% o más durante al menos 10 segundos.
- **Hipopnea**: disminución del FA del 30% o más durante al menos 10 segundos acompañado de una desaturación de oxígeno de al menos un 4%. Alternativamente, también se considera hipopnea cuando hay una reducción del flujo de aire del 50% o más durante al menos 10 segundos, con una desaturación de oxígeno del al menos el 3% o un *arousal* asociado.

Estos eventos se producen cuando las vías respiratorias superiores se colapsan o estrechan repetidamente, provocando pausas en la respiración o una diminución del flujo de aire. En consecuencia, los niveles de oxígeno en sangre disminuyen y se generan

episodios de hipoxia intermitente (Cuesta, 2005; Lloberes et al., 2011). Además, la desestructuración del sueño es otro factor determinante en la clínica (Lloberes et al., 2011). Durante la noche, los pacientes experimentan interrupciones frecuentes en el patrón normal del sueño, debido a los despertares causados por las apneas y la hipoxia. Estas interrupciones fragmentan el sueño, impidiendo al paciente alcanzar las fases de sueño profundo (Ho & Brass, 2011).

Estos procesos fisiológicos se traducen en la presencia de síntomas y signos que pueden manifestarse tanto durante el día como durante la noche (Lloberes et al., 2011). Los síntomas diurnos incluyen somnolencia excesiva, hipersomnia diurna, falta de energía, irritabilidad y disminución de la memoria y la concentración. Durante la noche, los pacientes pueden experimentar ronquidos fuertes, tos, despertares frecuentes y sensación de ahogo o asfixia (Eligulashvili & Pal'man, 1997). Además, una mala calidad del sueño podría estar relacionada con un mayor riesgo de demencia al verse comprometida la consolidación de la memoria y la remodelación sináptica (Pase et al., 2023).

Como se menciona anteriormente, uno de los principales efectos de la AOS es la hipoxemia, es decir, la disminución de los niveles de oxígeno en sangre. Además, se produce una acumulación de dióxido de carbono en la sangre, conocida como hipercapnia, teniendo consecuencias negativas en el sistema cardiovascular. Por tanto, es frecuente que los pacientes manifiesten hipertensión arterial, insuficiencia cardiaca, arritmias, cardiopatía isquémica y un mayor riesgo de padecer ictus (Lloberes et al., 2011).

1.1.3. DIAGNÓSTICO

El diagnóstico de la AOS se basa en la evaluación de los síntomas clínicos y los factores de riesgo, complementado por la realización de un estudio del sueño a través de una polisomnografía en un laboratorio especializado o mediante el uso de dispositivos portátiles diseñados para su utilización en el hogar (González Mangado et al., 2020; Ho & Brass, 2011).

En los que respecta a los síntomas clínicos, principalmente se estudia la presencia de ronquidos fuertes, pausas respiratorias durante el sueño, somnolencia diurna excesiva y/o hipertensión arterial (Eligulashvili & Pal'man, 1997). Asimismo, se deben considerar los factores de riesgo asociados, como la obesidad o la edad avanzada (Ho & Brass, 2011).

La evaluación clínica se complementa mediante la utilización de cuestionarios que ayudan a recopilar información relevante sobre lo síntomas del paciente y su impacto en la calidad de vida. Concretamente, la escala de somnolencia Epworth (*Epworth Sleepiness Scale,* ESS) ha sido adoptada a nivel mundial como método de cribado eficaz (Eguía & Cascante, 2007; Lloberes et al., 2011). Este cuestionario es gran utilidad para determinar la probabilidad de que el paciente experimente episodios de somnolencia en ocho situaciones distintas de la vida cotidiana (Ho & Brass, 2011).

Además de la evaluación clínica, se requiere una evaluación formal del sueño para confirmar el diagnóstico de la AOS. La polisomnografía (PSG) es considerada el *gold standard* y se realiza en un laboratorio especializado, donde se registran diversas variables

fisiológicas durante el periodo de sueño del paciente (Eligulashvili & Pal'man, 1997). Asimismo, existen también dispositivos portátiles de uso doméstico para una evaluación más accesible (Ho & Brass, 2011; Lloberes et al., 2011).

La evaluación del sueño mediante estas técnicas permite calcular el **índice de apnea-hipopnea** (*apnea-hypopnea index*, AHI), un parámetro utilizado para evaluar la gravedad de la AOS basado en el número total de apneas e hipopneas que ocurren por hora de sueño (Eligulashvili & Pal'man, 1997). En términos generales, se considera que una persona adulta padece AOS si presenta un AHI mayor o igual que 5 eventos por hora (e/h) de sueño, manifestando síntomas diurnos, o un AHI superior a 15 e/h, independientemente de los síntomas (Eguía & Cascante, 2007). Por otro lado, el AHI también se puede utilizar para estratificar la gravedad de la enfermedad en los siguientes niveles: no AOS (AHI < 5 e/h), AOS leve (5≤AHI< 15 e/h), AOS moderado (15≤AHI< 30 e/h), AOS grave (AHI≥30 e/h) (Eligulashvili & Pal'man, 1997; Ho & Brass, 2011).

1.2 POLISOMNOGRAFÍA

La PSG nocturna es ampliamente reconocida como el *gold standard* para el diagnóstico de pacientes con sospecha de AOS (Lloberes et al., 2011). Mediante esta técnica se registran de forma continua diversas señales neurofisiológicas y cardiorrespiratorias, lo que facilita la evaluación tanto de la cantidad como la calidad del sueño. Además, permite distinguir eventos respiratorios y su impacto en los sistemas corporales (Eguía & Cascante, 2007).

Durante la realización de la PSG, se registra la actividad electroencefalográfica (electroencefalografía ó EEG), que normalmente se lleva a cabo en al menos dos derivaciones centrales (C3 y C4) y, si es posible, en regiones occipitales (O1 y O2). Asimismo, se registra la actividad de los movimientos oculares (electrooculograma o EOG) y el tono muscular (electromiograma o EMG) en el mentón. Adicionalmente, se emplean electrodos específicos y sensores para capturar los movimientos de las extremidades inferiores y ayudar a determinar la posición corporal (Cuesta, 2005). Todas ellas se utilizan para cuantificar las fases del sueño y los *arousals* (Lloberes et al., 2011).

Con respecto al estudio de los parámetros respiratorios y cardiacos, se registran otras señales biomédicas como la saturación de oxígeno en sangre (SpO_2) mediante un pulsioxímetro, el esfuerzo respiratorio mediante bandas toracoabdominales, el flujo de aire oronasal mediante una cánula nasal y un termistor, los ronquidos y el electrocardiograma (Cuesta, 2005; Lloberes et al., 2011). Por otro lado, la realización de la polisomnografía se lleva a cabo durante la noche o el periodo habitual del sueño del paciente, con una duración mínima de 6,5 horas, incluyendo al menos 3 horas de sueño efectivo (Lloberes et al., 2011).

A pesar de que la PSG es un método eficaz, presenta ciertos inconvenientes que limitan su aplicabilidad (Lloberes et al., 2011). En primer lugar, es una técnica con un elevado coste económico, requiere una preparación laboriosa y no todos los centros médicos tienen acceso a esta tecnología. Asimismo, su análisis demanda una cantidad considerable de tiempo y mano de obra, aumentando el riesgo de cometer errores por parte de los profesionales. Por otro lado, es una prueba incómoda para el paciente, ya que es necesario

estar conectado a múltiples cables y sensores. Además, debido a la alta demanda de estos estudios, no siempre es posible aplicarla a todos los pacientes que la requieren, resultando en largas listas de espera (Mostafa et al., 2019). Estas limitaciones de la PSG, junto con la alta prevalencia de la enfermedad, han llevado a la comunidad científica a investigar el uso de pruebas simplificadas de detección de la AOS (Collop et al., 2011; del Campo et al., 2018).

1.3 PRUEBAS DE APNEA DEL SUEÑO EN CASA

La clasificación de los estudios del sueño por parte de la *American Academy of Sleep Medicine* (ASSM) se divide en cuatro tipos (Collop et al., 2011; Kapur et al., 2017):

- Tipo I: PSG convencional, la cual se realiza bajo la supervisión de un técnico de laboratorio de sueño.
- Tipo II: PSG portátil completa. Esta prueba se realiza de manera no supervisada con un equipo portátil y requiere un mínimo de siete canales por registro, incluyendo EEG, EOG, EMG submentoniano, ECG o frecuencia cardiaca, FA, esfuerzo respiratorio y SpO_2.
- Tipo III: poligrafía respiratoria (PR), que se enfoca en el registro de la ventilación (un mínimo de dos señales de movimiento respiratorio o una señal de movimiento respiratorio y el FA), además del ECG o frecuencia cardiaca, y SpO_2, utilizando un total de cuatro a siete canales de registro.
- Tipo IV: dispositivos de uno o dos canales, como la oximetría y/o la respiración. Todos los dispositivos que no cumplen los requisitos del tipo III se incluyen en este grupo.

Con el objetivo de afrontar las dificultades de la PSG, se han propuesto alternativas diagnósticas de la AOS conocidas como Pruebas de Apnea del Sueño en Casa o HSAT, por sus siglas en inglés (*Home Sleep Apnea Testing*) (Rundo & Downey, 2019). Esta modalidad de evaluación también se conoce como monitorización portátil y, como su nombre indica, se realiza mientras el paciente permanece en su hogar. Esto resulta beneficioso para aquellos que puedan sentirse incómodos al tener que pasar toda la noche en laboratorios del sueño. Además, son sistemas más compactos y menos engorrosos, y no requieren supervisión por parte de un técnico durante la realización de la prueba (Kapoor & Greenough, 2015).

Los dispositivos empleados en estudios de sueño domiciliarios (HSAT) pueden ser de Tipo III o Tipo IV y, por tanto, utilizan un menor número de sensores en total y no registran la actividad electroencefalográfica (EEG). Adicionalmente, estos dispositivos también pueden incluir sensores para evaluar la posición corporal, la frecuencia cardiaca o el pulso, y el movimiento como una medida sustitutiva del EEG (Rundo & Downey, 2019).

El EEG se utiliza para la detección de fases del sueño y *arousals*. En consecuencia, como la mayoría de HSAT no incluyen la monitorización de la actividad eléctrica cerebral, no son capaces de detectar eventos que únicamente se asocian a *arousals* (Kapoor & Greenough, 2015). Asimismo, en los HSAT no se puede distinguir entre estar

despierto o dormido, lo cual puede afectar a la estimación del AHI y, en consecuencia, el diagnóstico de la AOS. Por ejemplo, un paciente que se somete a una PSG puede pasar 10 horas en la cama, pero el EEG podría revelar que solamente durmió 5 horas. Si se registran un total de 100 eventos durante la prueba, teniendo en cuenta que hay 5 horas reales de sueño, equivaldría a un total de 20 eventos por hora. En cambio, al utilizar un HSAT en el paciente mencionado la gravedad se calcularía con el tiempo total de registro de 10 horas, lo que equivale a 10 eventos por hora. Por tanto, un HSAT tiende a subestimar la gravedad de la enfermedad. Esto es especialmente relevante en pacientes con AOS leve, ya que podrían obtener un resultado negativo (Kapoor & Greenough, 2015).

En consecuencia, la AASM recomienda el uso de las pruebas de sueño en el domicilio para evaluar la AOS en pacientes sin complicaciones, pero con un riesgo moderado o alto de padecerlo (Kapur et al., 2017). Además, no se recomienda el HSAT en pacientes mayores de 65 años, ya que no se ha investigado lo suficiente en esta población (Rundo & Downey, 2019).

Por otro lado, en los últimos años, ha habido un creciente interés en la pulsioximetría como alternativa a la PSG en el contexto del diagnóstico de la AOS debido a su simplicidad, fiabilidad y accesibilidad (del Campo et al., 2018). La oximetría de pulso se considera un dispositivo de Tipo IV para el diagnóstico de la AOS, ya que registra la SpO_2 y frecuencia cardiaca de manera no invasiva con un oxímetro de pulso, generalmente ubicado en el dedo, el dedo del pie o el lóbulo de la oreja del paciente (Netzer et al., 2001).

Dada la elevada prevalencia de la AOS, como se comentó anteriormente, los HSAT pueden resultar una herramienta muy útil en la atención primaria para evaluar a pacientes con sospecha de dicha patología, facilitar el diagnóstico temprano y anticipar el inicio del tratamiento, siempre y cuando se respeten los criterios de exclusión establecidos por la AASM. Además, son una alternativa diagnóstica más sencilla y económica en comparación con la PSG.

1.4 PULSIOXIMETRÍA

El contenido total de oxígeno (O_2) presente en la sangre se divide en dos componentes distintos. El primero corresponde al O_2 que se encuentra unido a la hemoglobina (Hb), representando aproximadamente un 97-98% del total presente en cuerpo humano. El segundo componente se refiere al O_2 que está disuelto en el plasma sanguíneo, aunque en una proporción muy inferior, alrededor del 2% al 3%. Por lo tanto, la Hb, presente en los glóbulos rojos, desempeña un papel esencial en la oxigenación sanguínea, siendo responsable de transportar el O_2 desde los pulmones hasta las células de los tejidos corporales (Nitzan et al., 2014).

La saturación de oxígeno en sangre (SaO_2) es una señal biomédica que proporciona información sobre la respiración. A diferencia de la señal de flujo aéreo, la forma de esta señal no sigue el patrón sinusoidal típico. Concretamente, en condiciones normales, la señal de SaO_2 es constante y plana (Cohen, 2006). La medición de la SaO_2 se realiza mediante el cálculo de la proporción entre la concentración de oxihemoglobina (HbO_2), que es la Hb asociada al oxígeno, y la concentración total de hemoglobina, que incluye tanto la hemoglobina oxigenada como la desoxihemoglobina (HHb) o hemoglobina libre

(Ecuación 1.1). Para ello se requiere extraer una muestra de sangre arterial, por lo que se trata de una prueba invasiva (Nitzan et al., 2014).

$$SaO_2 = \frac{[HbO_2]}{[HbO_2]+[HHb]} \quad (1.1)$$

No obstante, existe una alternativa para estimar la SaO_2 de forma transcutánea y continua: la técnica óptica de la pulsioximetría (PO) (Cohen, 2006). La pulsioximetría se basa en la detección de la señal fotopletismográfica (PPG), en la cual se distinguen componentes continuos (DC) y componentes pulsátiles (AC) superpuestos. El componente DC se compone de las influencias respiratorias, la actividad del sistema nervioso simpático y la termorregulación. Por su parte, el componente AC es originado por las fluctuaciones sincrónicas en el volumen sanguíneo que se generan como resultado de la actividad cardíaca, estando vinculado a las fases sistólica y diastólica del ciclo. La fase sistólica tiene su inicio en un punto mínimo y finaliza al alcanzar el pico sistólico de la onda de pulso. A su vez, en esta onda de pulso se puede apreciar otro punto mínimo que marca el final de la fase diastólica (Chan et al., 2013; Ghamari, 2018).

La señal de PPG es provechada para evaluar las variaciones en los intervalos temporales en los latidos cardiacos (intervalo de pico a pico o P-P), como se ilustra en la Figura 1.1 (Ghamari, 2018). Estas variaciones pueden deberse a diversos factores, tales como la edad del individuo, sus condiciones cardiacas y su estado físico. Finalmente, a partir del intervalo P-P, se deriva la señal de PR (Chan et al., 2013; Ghamari, 2018).

Por otro lado, la SaO_2 se obtiene considerando exclusivamente las variaciones en el tiempo de la absorbancia causada por la sangre arterial pulsante utilizando longitudes de onda rojas e infrarrojas (Cohen, 2006). Esto se debe a que la oxihemoglobina y desoxihemoglobina presentan diferentes patrones de absorción de la luz roja y la luz cercana al infrarrojo del espectro de longitudes. En la Figura 1.2 se puede observar que la HbO_2 absorbe menor luz roja y mayor cantidad de luz infrarroja, mientras que la HHb absorbe más luz roja (Madhan Mohan et al., 2016; Nitzan et al., 2014). Al valor de la SaO_2 estimado mediante esta técnica se le conoce como saturación de oxígeno en sangre periférica (SpO_2) (Madhan Mohan et al., 2016).

Los pulsioxímetros aprovechan esta característica emitiendo luz en dos longitudes de onda diferentes: roja a 660 nm e infrarroja a 940 nm. Posteriormente, la luz atraviesa los diversos tejidos y finalmente es capturada por un fotodiodo localizado en el lado opuesto del sensor, como se ilustra en la Figura 1.3 (Madhan Mohan et al., 2016).

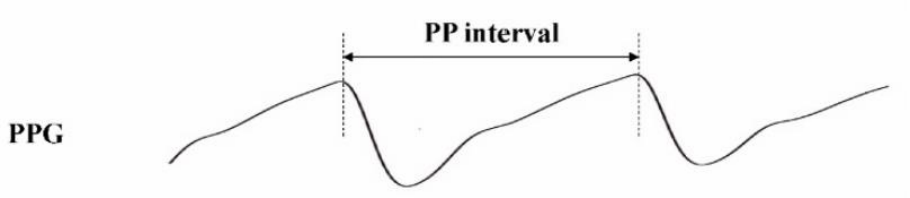

Figura 1.1. Intervalo P-P de una señal de PPG (Ghamari, 2018).

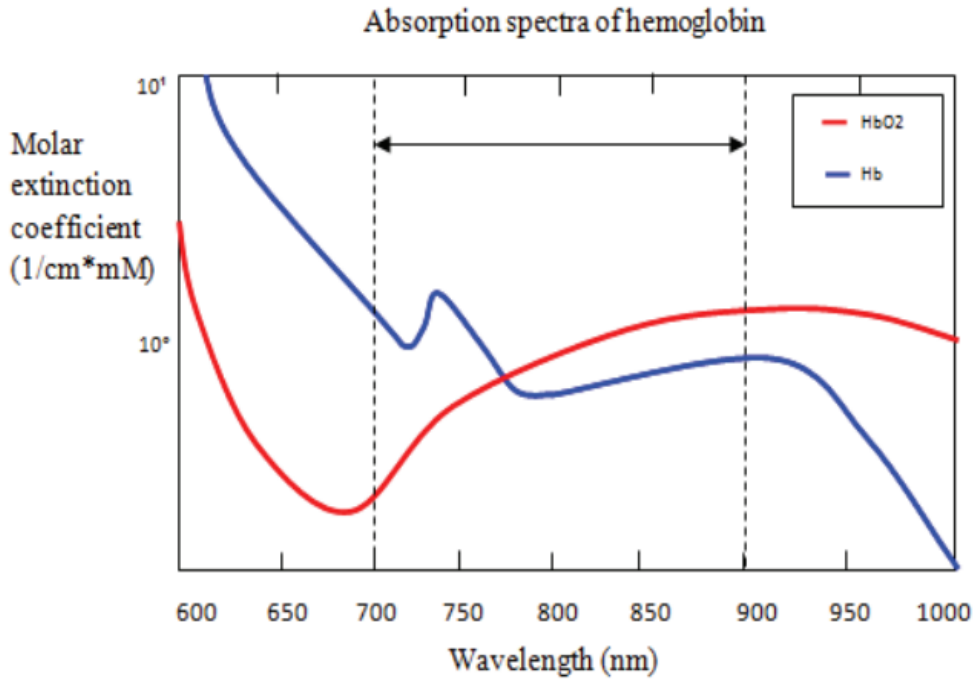

Figura 1.2. Espectro de absorción de la hemoglobina oxigenada y desoxigenada (Madhan Mohan et al., 2016).

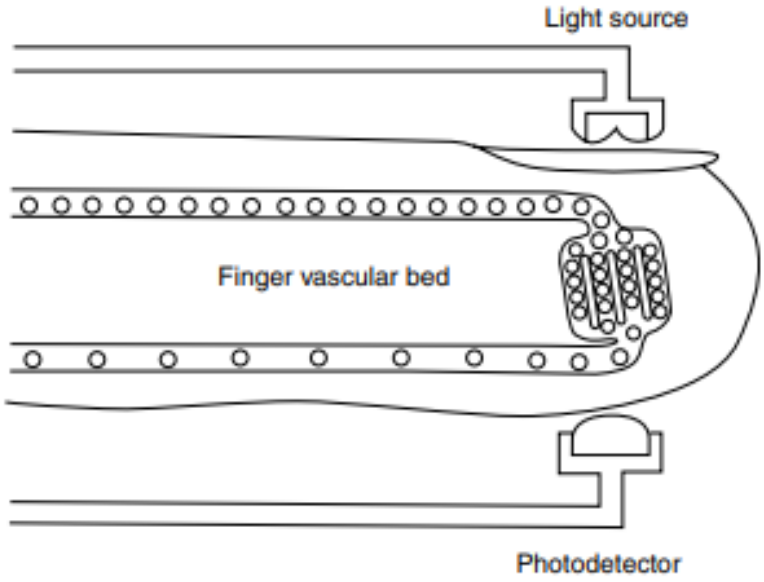

Figura 1.3. Configuración típica del pulsioxímetro en el dedo índice (Cohen, 2006).

La transmisión de la luz en cada longitud de onda está influenciada por el grosor, el color y la estructura del material que atraviesa. Para cuantificar la absorbancia, se aplica la Ley de Beer-Lambert, que establece que la absorbancia es directamente proporcional a la concentración y al camino óptico del material absorbente. De este modo, la pulsioximetría permite calcular de manera precisa la saturación de oxígeno en sangre. Para realizar esta medición, típicamente se emplean el dedo índice, el lóbulo de la oreja o el dedo gordo del pie en el caso de los niños (Cohen, 2006).

La medición de la saturación de oxígeno se expresa en porcentaje (%), ya que se trata de una relación entre la oxihemoglobina y el total de hemoglobina en la sangre. Por lo

general, esta señal se mantiene en niveles constantes en torno al 96,5% ± 1,5%, lo cual se considera un rango de normalidad en individuos sanos sin ninguna condición patológica (Netzer et al., 2001). De esta manera, valores que oscilan entre el 95% y el 100% se consideran dentro de los rangos normales, mientras que valores inferiores al 90% reflejan situaciones de hipoxemia.

La pulsioximetría es una técnica muy empleada en estudios del sueño para llevar a cabo monitorizaciones de larga duración y registrar eventos respiratorios durante este periodo (Netzer et al., 2001). Esta herramienta permite identificar y cuantificar la presencia de episodios de hipoxemia intermitente, caracterizados por desaturaciones transitorias pero recurrentes a lo largo del tiempo. Es importante resaltar que los mecanismos fisiológicos que se producen desde el cese de la respiración hasta la reducción de los niveles de oxígeno en sangre no son instantáneos, sino que tienen lugar aproximadamente entre 10 y 30 segundos después (Otero et al., 2012). En otras palabras, existe un retardo variable entre el comienzo del evento respiratorio registrado en la señal de FA y su correspondiente desaturación detectada en la señal de SpO_2 (Kulkas et al., 2013).

Esta técnica ofrece numerosas ventajas, como la capacidad de controlar la oxigenación sanguínea de forma continua, segura, efectiva y no invasiva sin requerir calibración previa antes de cada uso (Cohen, 2006). En consecuencia, la pulsioximetría se ha convertido en una herramienta de gran utilidad en los estudios domiciliarios del sueño (HSAT) para medir la SaO_2 en tiempo real y monitorizar de forma indirecta la respiración.

1.5 DEEP LEARNING

La Inteligencia Artificial (IA) es un campo enfocado a desarrollar máquinas que sean capaces de realizar funciones normalmente asociadas al cerebro humano, como el aprendizaje y la resolución de problemas. El aprendizaje es un componente fundamental en las máquinas dotadas de inteligencia artificial. Por lo tanto, el aprendizaje automático o *machine learning* (ML) se considera un campo específico dentro de la IA. Estos algoritmos buscan que las máquinas sean capaces de aprender y mejorar su rendimiento a través de la experiencia sin haber sido programadas explícitamente para ello (Shinde & Shah, 2018).

Las técnicas convencionales de ML requieren una cuidadosa etapa de extracción de características, que transforme los datos sin procesar (*raw data*) en vectores de características. Esta etapa es esencial para permitir que el sistema de aprendizaje, generalmente un clasificador, pueda detectar y clasificar patrones en la entrada. Sin embargo, este enfoque implica una gran dependencia con el conocimiento del experto para diseñar manualmente las características más relevantes para un problema específico (Lecun et al., 2015).

El aprendizaje de representaciones o *representation learning* es un conjunto de métodos que permite a los sistemas aprender automáticamente de las características relevantes a partir de los datos sin procesar. En particular, el aprendizaje profundo, también denominado *deep learning*, destaca como un enfoque de *representation learning* que emplea una estructura en cascada compuesta por múltiples capas de unidades de

procesamiento no lineales encargadas de extraer características y realizar transformaciones a los datos de entrada (Lecun et al., 2015).

En las capas inferiores, las unidades de procesamiento aprenden a reconocer características simples y a medida que se avanza hacia capas superiores la representación es cada vez más abstracta, permitiendo aprender estructuras y patrones altamente complejos. Esto resulta especialmente útil en tareas de clasificación, donde las capas superiores retienen las características más relevantes de la entrada, al mismo tiempo que suprimen los rasgos que proporcionan menor información. Además, esta arquitectura jerárquica es adecuada para extraer conocimiento a partir de grandes volúmenes de datos (Lecun et al., 2015; Shinde & Shah, 2018).

En el campo de reconocimiento de voz, el aprendizaje profundo ha logrado avances significativos en tareas como la detección automática del habla y la transcripción de voz a texto. Por otro lado, ha resultado ser muy efectivo en el ámbito de visión por computadora en aplicaciones como clasificación de imágenes, detección de objetos y reconocimiento facial. Otra de sus principales aplicaciones es el procesamiento de lenguaje natural mejorando el rendimiento en la clasificación de textos, en el análisis de sentimientos o en la traducción automática (Lecun et al., 2015; Najafabadi et al., 2015).

Finalmente, las técnicas de *deep learning* han demostrado su utilidad para el análisis y clasificación de series temporales, donde se requiere clasificar patrones en datos secuenciales a lo largo del tiempo (Ismail Fawaz et al., 2019). Concretamente, en este trabajo se va a emplear para la detección automática de eventos respiratorios a partir de las señales de SpO_2 y PR.

1.5.1 EXPLAINABLE ARTIFICIAL INTELLIGENCE

Desde sus inicios, las aplicaciones de IA son altamente demandadas en el ámbito de la salud, especialmente en la ayuda a la decisión clínica. Por ejemplo, numerosos modelos de DL son desarrollados para tareas específicas como la interpretación de electrocardiogramas, el diagnóstico de enfermedades o la selección de tratamientos adecuados. No obstante, estos modelos son percibidos como cajas negras, sin ofrecer una explicación transparente de cómo se generan sus predicciones. En consecuencia, surge la necesidad de proporcionar una explicación clara de los resultados generados por dichos modelos (Loh et al., 2022).

En este contexto surge la *Explainable Artificial Intelligence* (XAI), definida como el conjunto de características que explican todo el proceso que sigue el modelo hasta llegar a la predicción final (Barredo Arrieta et al., 2020). En otras palabras, XAI permite a los usuarios comprender el funcionamiento interno del modelo detrás de cada predicción. Por ejemplo, en la tarea de clasificación de imágenes médicas, XAI tiene la capacidad de explicar cómo opera el modelo, destacando las partes más influyentes de la imagen que han conducido a la predicción. Asimismo, también ha resultado ser efectiva para identificar patrones de interés en señales unidimensionales (Loh et al., 2022). Por lo tanto, este enfoque es necesario para fomentar la confianza en la IA, y, sobre todo, en los modelos de DL, entre los profesionales de la salud (Loh et al., 2022).

Por otro lado, *Gradient-weighted Class Activation Mapping* (Grad-CAM) es una técnica de XAI muy popular entre las arquitecturas de DL basadas en capas convolucionales (Selvaraju et al., 2020). Su diseño está orientado específicamente a la identificación de características discriminativas en las capas convolucionales para las predicciones del modelo. En consecuencia, ha demostrado un gran potencial en la interpretación de imágenes médicas. No obstante, su aplicabilidad también se extiende al ámbito de la toma de decisiones médicas. Asimismo, ha resultado ser efectiva al aplicarse a señales unidimensionales, como el electrocardiograma y el electromiograma, para identificar el momento temporal en el cual se presenta alguna anomalía (Loh et al., 2022).

1.6 HIPÓTESIS Y OBJETIVOS

Para realizar este trabajo de fin de grado (TFG), se plantea la hipótesis de que los métodos de *deep learning* pueden detectar de forma automática los diferentes eventos respiratorios durante el sueño partir de señales de pulsioximetría. Continuando con dicha hipótesis, el objetivo general de este trabajo consiste en desarrollar y evaluar diversos modelos de *deep learning* para clasificar automáticamente los eventos de apnea e hipopnea a partir de las señales de SpO_2 y PR procedentes de la pulsioximetría. Para lograr este objetivo, se llevará a cabo un análisis de los registros nocturnos de dichas señales en adultos incluidos en la base de datos pública del estudio *Multi-Ethnic Study of Atherosclerosis* (MESA), disponible en https://sleepdata.org/datasets/mesa.

Una vez definido el objetivo general, se detallan los objetivos específicos de este trabajo, que son necesarios para alcanzar el objetivo general:

1. Realizar una revisión bibliográfica en revistas científicas enfocadas en técnicas de *deep learning* aplicadas para la detección automática de eventos de apnea e hipopnea.
2. Seleccionar la técnica de *deep learning* más apropiada para la detección de estos eventos respiratorios y proceder a su implementación.
3. Aplicar la técnica seleccionada a las señales de SpO_2 y PR de la base de datos disponible de MESA, evaluando su precisión en la detección de eventos de apnea e hipopnea, así como la estimación del AHI.
4. Analizar en profundidad el algoritmo implementado para identificar posibles aspectos a mejorar.
5. Comprender el proceso de toma de decisión del algoritmo para la detección de apneas e hipopneas mediante enfoques de XAI.
6. Extraer conclusiones a partir de los resultados obtenidos, así como sugerir posibles líneas de investigación futuras.

1.7 PLANIFICACIÓN Y ESTRUCTURA DEL TFG

1.7.1 PLAN DE TRABAJO

Con el fin de alcanzar el objetivo general y los diferentes objetivos específicos, se ha implementado un plan de trabajo que consta de tres fases principales:

Fase 1: formación básica y contextualización del problema

- Realización del curso de Coursera '*Deep Learning Specialization'*.
- Búsqueda bibliográfica de los métodos publicados para la detección automática de eventos de apnea e hipopnea a partir de señales cardiorrespiratorias.
- Adquisición de conocimientos en el lenguaje de programación Python y familiarización con el manejo de las librerías necesarias para implementar redes neuronales convolucionales (CNN).
- Análisis y preprocesado de la base de datos de señales de SpO_2 y PR disponible: MESA.

Fase 2: diseño y aplicación del método de *deep learning*

- Diseño de la metodología a implementar.
- Aplicación del método de *deep learning* seleccionado a la base de datos de MESA con el fin de evaluar la precisión en la detección de eventos de apnea e hipopnea.
- Obtención de diferentes métricas de rendimiento para la detección de eventos de apnea e hipopnea y la estimación del AHI.

Fase 3: conclusiones e informe final

- Análisis de los resultados en términos de rendimiento y comparación con los publicados en estudios previos.
- Interpretación de las arquitecturas CNN mediante métodos de XAI: identificación de patrones de las señales de SpO_2 y PR relacionados con apneas e hipopneas.
- Extracción de conclusiones a partir de todos los modelos realizados e identificación de las principales limitaciones.
- Redacción del informe final.

1.7.2 ESTRUCTURA DE LA MEMORIA DEL TFG

El TFG se divide en siete capítulos: introducción, detección automática de eventos de apnea e hipopnea, sujetos y señales, metodología, resultados, discusión, conclusiones y líneas futuras.

El Capítulo 1: 'Introducción' aborda la AOS, un trastorno respiratorio común que se caracteriza por eventos recurrentes de apnea e hipopnea durante el sueño. Asimismo, se explora la PSG como principal técnica diagnóstica, se describen las pruebas de apnea del sueño domiciliarias (HSAT), particularizando en la pulsioximetría, una técnica de medición no invasiva para monitorizar los niveles de oxígeno en sangre. Además, se introduce el concepto de *deep learning* en el ámbito del análisis de señales biomédicas,

así como el de XAI. Finalmente, se presentan las hipótesis y los objetivos de este trabajo de investigación.

El Capítulo 2: 'Detección automática de eventos de apnea e hipopnea' se centra en el desafío de clasificar de forma automática los eventos de apnea e hipopnea. Para ello, se exploran diferentes enfoques basados en *deep learning* implementados en la comunidad científica y se comparan para seleccionar el método más adecuado para desarrollar en este trabajo.

En el Capítulo 3: 'Sujetos y señales' se describe la población bajo estudio correspondiente a la base de datos MESA. Además, se detallan las características de las señales de pulsioximetría utilizadas en este trabajo y se explica la metodología empleada para dividir las diferentes instancias en conjuntos de entrenamiento, validación y test.

En el Capítulo 4: 'Metodología' se examina la estructura general de las arquitecturas CNN y se resaltan sus ventajas en comparación con las redes neuronales artificiales (ANN) convencionales. Asimismo, se introduce el concepto de XAI. Finalmente, se exponen los estadísticos utilizados para evaluar el rendimiento de la clasificación de eventos de apnea e hipopnea, así como la estimación del AHI.

En el Capítulo 5: 'Resultados' se presentan los resultados obtenidos empleando la señal de SpO_2 y la señal de PR de forma individual, así como ambas combinadas. Además, se presentan los resultados tanto para segmentos de 30 segundos como para segmentos de 60 segundos, considerando casos sin adyacencia y con adyacencia. Finalmente, se presentan los mapas de calor de los modelos CNN obtenidos mediante XAI.

En el Capítulo 6: 'Discusión' se analizan y discuten los resultados obtenidos en la clasificación de eventos de apnea e hipopnea, así como en la estimación del AHI, y se comparan con los obtenidos en otros estudios relacionados. Asimismo, se proporciona una interpretación de los patrones de la señal de SpO_2 y PR relacionados con eventos de apnea e hipopnea. Finalmente, se identifican las principales limitaciones del trabajo.

Por último, en el Capítulo 7: 'Conclusiones y líneas futuras' se comentan las conclusiones, resaltando las contribuciones realizadas en la detección automática de eventos de apnea e hipopnea mediante técnicas de *deep learning*. Finalmente, se plantean posibles líneas de investigación futuras para mejorar los resultados y abordar las limitaciones identificadas.

CAPÍTULO 2: DETECCIÓN AUTOMÁTICA DE EVENTOS DE APNEA E HIPOPNEA

La AOS es un trastorno respiratorio muy frecuente que afecta a casi mil millones de personas en todo el mundo (Benjafield et al., 2019). En consecuencia, es de vital importancia conseguir métodos eficientes de detección temprana y precisa para prevenir las posibles complicaciones en estos pacientes y mejorar su calidad de vida.

En los últimos años, el empleo de técnicas de *deep learning* ha mostrado un gran potencial en el diagnóstico de la AOS. En este contexto, se ha llevado a cabo una revisión bibliográfica con el fin de analizar el estado del arte en la detección automática de eventos de apnea e hipopnea, recopilando un total de veinte artículos publicados a partir de 2017 que emplearon señales cardiorrespiratorias. A través de esta revisión, se busca obtener una visión actualizada de las investigaciones más recientes que empleen técnicas de *deep learning* para detectar eventos de apnea e hipopnea y mejorar el diagnóstico de la AOS. Además, va a permitir comparar entre las diferentes metodologías para identificar cuál de los métodos es el más adecuado para implementar en este TFG.

2.1 ESTUDIOS PREVIOS

Con el objetivo de presentar de manera más clara y fácilmente interpretable la información recopilada, se ha realizado una clasificación en dos tablas. La 'Tabla 2.1' agrupa todos los estudios que emplean la señal de ECG, mientras que en la 'Tabla 2.2' se encuentran los estudios que utilizaron otras señales diferentes, como SpO_2 o FA. Ambas tablas contienen detalles sobre las bases de datos utilizadas, la metodología implementada y los resultados obtenidos por los distintos investigadores. Dentro de la metodología, se especifica la forma en que los segmentos fueron etiquetados, el tamaño del segmento de entrada y el método específico de *deep learning empleado*. Además, se indica si se llevó a cabo o no la estimación del AHI para realizar una clasificación por sujeto.

2.1.1 ARQUITECTURAS DE DEEP LEARNING

Entre los diversos enfoques analizados dentro del campo del *deep learning*, se han mencionado varios artículos que utilizan arquitecturas CNN (Cen et al., 2018; Chang et al., 2020; Choi et al., 2018a; Dey et al., 2018; Haidar et al., 2018; Leino et al., 2021; Mostafa et al., 2020; Nasifoglu & Erogul, 2021; Sharan et al., 2020; Urtnasan et al., 2018; T. Wang et al., 2019; X. Wang et al., 2020). En todos ellos se pueden observar características comunes en cuanto al diseño de sus modelos.

AUTOR	SEÑAL	BASE DE DATOS	METODOLOGÍA				RESULTADOS
			MÉTODO	TAMAÑO	ETIQUETA	AHI	
(Urtnasan et al., 2018)	ECG	Privada	CNN	10s	N/A/H	No	**Segmento**: *F1-Score* (0.93), *precision* (0.87), *recall* (0.87), Acc (0.908)
(Dey et al., 2018)	ECG	Apnea-ECG	CNN	60s	N/A	No	**Segmento**: Acc (98.91%), Se (97.82), Sp (99.20), PPV (99.06%), NPV (98.14%)
(T. Wang et al., 2019)	ECG*	Apnea-ECG//UCD	Le-Net	60s	N/A	Si	**Segmento**: Se (83.1%), Sp (90.3%), Acc (87.6%), AUC (0.95). **Sujeto**: Acc (97.1), Se (100%), Sp (91.7%), AUC (0.996), Corr (0.943)
(Chang et al., 2020)	ECG	Apnea-ECG	CNN	60s	N/A	Si	**Segmento**: Acc (87.9%), Se (81.1%), Sp (92%). **Sujeto**: Acc (97.1%), Se (95.7%), Sp (100%)
(Nasifoglu & Erogul, 2021)	ECG**	Apnea-ECG//Privada	CNN	30s	N/A	Si	**Sujeto**: Acc (91.93%)
(J. Zhang et al., 2021)	ECG	Apnea-ECG	CNN-LSTM	10s con *overlap*.	N/A	No	**Segmento**: Kp (0.92), Acc (96.1%), Se (96.1%), Sp (96.2%), PPV (97.6%), NPN (93.8%)
(X. Wang et al., 2020)	ECG*	Apnea-ECG	CNN	-	N/A	No	**Segmento**: Acc (97.8%), Se (100%), Sp (93%)
(Almutairi et al., 2021a)	ECG*	Apnea-ECG	CNN//CNN+LSTM//CNN+GRU	60s	N/A	No	**Segmento**: Acc (89.11%), Se (89.91%), Sp (87.78%), *F1-score* (91.41%) para el modelo CNN+LSTM
(Qin & Liu, 2022)	ECG	Apnea-ECG//Privada	CNN-GRU	60s	N/A	Si	**Segmento**: Acc (91.1%), Se (88.9%), Sp (92.4%), F1 (0.883). **Sujeto**: Acc (100%), Se (100%), Sp (100%), F1 (1.000)
(Sharan et al., 2020)	ECG*	Apnea-ECG	CNN	60s	N/A	No	**Segmento**: Acc (88.23%), Se (82.74%), Sp (91.62%), AUC (0.9453)
(Erdenebayar et al., 2019)	ECG	Privada	DNN, CNN (1D-2D), RNN, GRU, LSTM	10s	N/A/H	No	**Segmento**: 1D CNN: Acc (96.3%), Se (96%), Sp (96%), Kappa (0.92). GRU: Acc (95%), Se (95%), Sp (96%), Kappa (0.91%)
(Liu et al., 2023)	ECG	Apnea-ECG	CNN+Transformer	3 min	N/A	Si	**Segmento**: Acc (88.2%), Se (78.5%), Sp (94.1%), Pre (89%). **Sujeto**: Acc (100%) para 3 min

*Tabla 2.1 Resumen revisión bibliográfica con señales de ECG. Anotaciones: *Intervalos RR y amplitudes del complejo QRS; **Escalogramas y espectogramas. Siglas: normal (N), apnea (A), hipopnea (H), accuracy (Acc), sensibilidad (Se), especificidad (Sp), deep neaural network (DNN), convolutional neural network (CNN), recurrent neural network (RNN), gated recurrent unit (GRU), long-short term memory (LSTM).*

AUTOR	SEÑAL	BASE DE DATOS	METODOLOGÍA				RESULTADOS
			MÉTODO	TAMAÑO	ETIQUETA	AHI	
(Choi et al., 2018a)	Presión nasal	MESA//Privada	CNN	10s con *overlap* cada 1s.	N/A	Si	**Segmento**: Kappa (0.82), Se (81.1%), Sp (98.5%), Acc (96.6%), PPV (87%), NPV (97,7%)
(Cen et al., 2018)	SpO_2, flujo oronasal, movimientos torácicos y abdominales	UCD	CNN	1s	N/A/H	No	**Segmento**: Acc global (79.61%)
(Haidar et al., 2018)	Flujo nasal, movimientos torácicos y abdominales	MESA	CNN	30s	N/A/H	No	**Segmento**: Acc (83.5%), *recall* (83.4), *precision* (83.4), *F1-score* (83.4) para los tres canales combinados
(Leino et al., 2021)	SpO_2	Privada//Privada	CNN	10 min con 98% de *overlap*.	N/A	Si	**Sujeto**: Test 1: REI (0.982). Test 2: REI (0.972)
(Pathinarupothi et al., 2017)	SpO_2//IHR	Privada	LSTM	60s	N/A	No	**Segmento**: Acc (95.5%), *Precision* (99.2%), *Recall* (92.9%), AUC (0.98)
(Huttunen et al., 2023)	PPG y SpO_2***	Privada	U-time	1s	N/A/H	Si	**Sujeto**: ICC (0.946), kappa (0.54)
(Nikkonen et al., 2021)	SpO_2, *thermistor-airflow, nasal, pressure-airflow, thorax respiratory effort*	Privada	LSTM	30s con 28s de *overlap*.	N/A/H	Si	**Segmento**: Kappa (0.728), ICC (0.985). **Sujeto**: ICC (0.985)
(Mostafa et al., 2020)	SpO_2	HuGCDN2008// Apnea-ECG//UCD	CNN	1 min// 3 min// 5 min	N/A	No	**Segmento**: AED 1 min: Acc(94.24%), Se (92.04%), Sp(95.78%); AED 3 min: Acc(93.93%), Se (89.87%), Sp(96.78%); UCD 1 min: Acc(84.85), Se(58.32%), Sp(93.32%); UCD 3 min: Acc(85,79%), Se(60.38%), Sp(93.9%)

*Tabla 2.2. Resumen revisión bibliográfica con diferentes señales. Anotaciones: *** La señal de SpO_2 y PPG se utiliza en el modelo 1. Siglas: normal (N), apnea (A), hipopnea (H), accuracy (acc), sensibilidad (Se), especificidad (Sp), ICC intraclass correlation coeficient (ICC), índice de eventos respiratorios (REI), area under curve (AUC), convolutional neural network (CNN), gated recurrent unit (GRU), long-short term memory (LSTM), apnea-ECG database (AED).*

La mayoría de los estudios emplean CNNs, que han demostrado su utilidad para analizar datos de señales biomédicas durante el sueño mediante una arquitectura multicapa con pesos compartidos, conexiones dispersas y elementos de reducción de dimensionalidad (Cen et al., 2018; Chang et al., 2020; Choi et al., 2018a; Dey et al., 2018; Haidar et al., 2018; Leino et al., 2021; Mostafa et al., 2020; Nasifoglu & Erogul, 2021; Sharan et al., 2020; Urtnasan et al., 2018; T. Wang et al., 2019; X. Wang et al., 2020). En primer lugar, todos los artículos utilizan un determinado número de bloques convolucionales, que constan de capas convolucionales, activación y *pooling* para extraer las características más relevantes de la señal. Asimismo, muchos estudios incorporan capas *batch normalization* en los bloques convolucionales. Por otra parte, el número de capas convolucionales varía entre los artículos, aunque, en general, no emplean un número elevado, ya que las redes tienden aprender rápidamente y la complejidad y el coste

computacional aumentan con el número de capas. En segundo lugar, la mayoría utilizan la función de activación ReLU (*Rectified Linear Unit*) en todas las capas, exceptuando la capa de salida. Concretamente, en la capa de salida se utiliza una función de activación *softmax* para obtener la probabilidad de cada clase. Por otro lado, el método de optimización Adam ha sido una elección común entre los estudios para entrenar las CNN (Cen et al., 2018; Chang et al., 2020; Choi et al., 2018a; Dey et al., 2018; Haidar et al., 2018; Leino et al., 2021; Mostafa et al., 2020; Nasifoglu & Erogul, 2021; Sharan et al., 2020; Urtnasan et al., 2018; T. Wang et al., 2019; X. Wang et al., 2020) .

Dentro esta estrategia de usar CNNs, Wang et al. (2019) proponen un enfoque interesante utilizando una arquitectura denominada LeNet-5 para aprender automáticamente características del ECG. Además, introducen una modificación adicional que consiste en la incorporación de segmentos adyacentes, lo que permite capturar relaciones temporales más complejas para mejorar la detección de eventos. Por otra parte, Nasifoglu et al. (2021) exploran la técnica de *Transfer Learning*, en la cual se aprovecha el conocimiento previamente adquirido por redes ya entrenadas en grandes conjuntos de datos. En este caso, utilizan como base las arquitecturas GoogleNet, AlexNet y ResNet18, adaptándolas para el problema de detección de eventos de apnea e hipopnea.

Como alternativa a las CNN, los artículos más recientes incorporan redes neuronales recurrentes (RNN), que permiten almacenar información de las secuencias previas y mejorar la capacidad del modelo para analizar patrones temporales en los datos de entrada, lo que contribuye a una detección más precisa de los eventos respiratorios (Erdenebayar et al., 2019; Nikkonen et al., 2021; Pathinarupothi et al., 2017). Concretamente, las RNN suelen utilizar capas *gated recurrent unit* (GRU) o *long-short term memory* (LSTM) (Erdenebayar et al., 2019; Nikkonen et al., 2021; Pathinarupothi et al., 2017). La principal diferencia entre ambas es que las GRUs tienen menos puertas o *gates*, lo que las hace más eficientes computacionalmente al tener que realizar menos cálculos. Por tanto, son una variante simplificada de las LSTM (Shiri et al., 2023). Asimismo, algunos estudios también combinan CNN y RNN para aprovechar sus ventajas (Almutairi et al., 2021a; Qin & Liu, 2022; J. Zhang et al., 2021). Por ejemplo, Qin & Liu et al. (2022) emplean un modelo CNN con una GRU bidireccional (BiGRU) para introducir dependencias temporales largas, tanto pasadas como futuras, que tengan que ver con el patrón de transición normal-apnea. Además, en la CNN utilizan *inception blocks* para mejorar la capacidad de aprendizaje de características y abordan el problema de desequilibrio de clases mediante el método de muestreo sintético adaptativo (ADASYN).

Además, es importante mencionar varias investigaciones que exploran diferentes enfoques de *deep learning* con el fin de encontrar el método óptimo para la detección automática de eventos de apnea del sueño. Por ejemplo, Almutairi et al. (2021a) evalúan tres modelos CNN, CNN+LSTM y CNN+GRU y destacan la eficacia del modelo CNN+LSTM para la detección de AOS. De forma similar, Erdenebayar et al. (2019) desarrollan y evalúan seis enfoques, incluyendo redes neuronales profundas (DNN), redes neuronales convolucionales unidimensionales y bidimensionales (CNN 1D y CNN 2D), RNN, GRU y LSTM. Tras esta investigación concluyeron que tanto las CNN 1D como las GRU presentan un gran potencial como herramientas automáticas para la detección de AOS, obteniendo los mejores rendimientos.

Otra arquitectura diferente es la utilizada por Huttunen et al. (2023), que consiste en una U-Net, es decir, una estructura *encoder-decoder* con bloques de convoluciones consecutivas y *skip connections* desde las capas *encoder* hasta las *decoder*. Se probaron tres combinaciones diferentes de señales de entrada en la misma arquitectura (modelo 1: fotopletismograma (PPG) y SpO_2; modelo 2: PPG, SpO_2 y presión nasal; modelo 3: SpO_2, presión nasal, EEG, termopar oronasal y cinturón respiratorio). Los resultados mostraron que el modelo 1 alcanzó un rendimiento comparable con los modelos 2 y 3 para la estimación del AHI. Sin embargo, la discriminación entre apnea e hipopnea era mucho peor, ya que esta aumentaba conforme se iban añadiendo señales.

Como se puede apreciar, hay distintos enfoques que involucran RNN e incluso la combinación de RNN con CNN. Sin embargo, es importante destacar una de las últimas tendencias emergentes del campo del *deep learning*: los *Transformers*. En el artículo recientemente publicado por Liu et al. (2023) implementan una estructura basada en CNN + *Transformer* para la detección automática de la AOS. El *Transformer* consiste en una estructura *encoder-decoder*. En primer lugar, el *encoder* toma una secuencia de entrada y la procesa en paralelo mediante unas capas denominadas *attention layers* y *feed-forward layers*, capturando las relaciones entre todas las partes de la secuencia de entrada. A su vez, el *decoder* emplea *global average pooling* (GAP) y *feed-forward neural networks* (FNN), junto con una capa de clasificación *softmax* para centrarse en las partes más relevantes del *encoder*. La arquitectura utilizada por Liu et al. (2023) consta de dos partes principales: la primera parte consiste en una CNN para aprender las características representativas de las señales de ECG, mientras que en la segunda parte se emplea el *Tranformer* para capturar el contexto temporal global y llevar a cabo la tarea de clasificación. Finalmente, en esta investigación también probaron el rendimiento del modelo sustituyendo el *Transformer* por LSTM, BiLSTM y GRU para segmentos de entrada de 1 y 3 minutos. Analizando los resultados, se puede observar que la estrategia con mejor rendimiento es utilizar una CNN seguida de un *Transformer* con tamaños de entrada de 3 minutos.

2.1.2 SEÑALES DE ENTRADA

Por otro lado, durante esta revisión bibliográfica se ha encontrado que la mayoría de las investigaciones se basan en el uso de la señal de ECG de una sola derivación como entrada del modelo para llevar a cabo la tarea de detección automática de eventos de apnea e hipopnea (Almutairi et al., 2021a; Chang et al., 2020; Dey et al., 2018; Erdenebayar et al., 2019; Liu et al., 2023; Nasifoglu & Erogul, 2021; Qin & Liu, 2022; Sharan et al., 2020; Urtnasan et al., 2018; T. Wang et al., 2019; X. Wang et al., 2020; J. Zhang et al., 2021). Además, en algunos estudios utilizan un solo canal de entrada con la señal de SpO_2 (Huttunen et al., 2023; Mostafa et al., 2020; Pathinarupothi et al., 2017) o incluso la presión nasal, alcanzando unos buenos resultados en términos de exactitud (Choi et al., 2018a).

Al analizar los resultados obtenidos en sus investigaciones, se aprecia que la señal de ECG de una sola derivación es la más efectiva, mostrando los mejores resultados en términos de exactitud tanto en la tarea de clasificación por segmento como en la tarea de clasificación por sujeto. No obstante, es importante señalar que estos resultados tan

elevados podrían estar influidos por la disponibilidad pública de la base de datos Apnea-ECG, la cual ha sido utilizada en la mayoría de los artículos que emplean esta señal. Por otra parte, entre las señales restantes, la señal de SpO_2 ha demostrado un buen rendimiento en la detección de eventos de apnea e hipopnea, así como en la estimación del AHI, lo que la convierte en una alternativa valiosa.

Asimismo, en varios artículos optan por combinar múltiples canales de entrada para obtener información complementaria y mejorar la exactitud del clasificador de eventos. Para ello, generalmente se incluyen la señal de flujo aéreo y los movimientos torácicos y abdominales junto con las señales previamente mencionadas, demostrando ser beneficiosos en la tarea de detección de eventos respiratorios (Cen et al., 2018; Haidar et al., 2018; Huttunen et al., 2023; Nikkonen et al., 2021). Concretamente, Huttunen et al. (2023) concluyeron que, al agregar señales adicionales a la señal de SpO_2, se mejoraba considerablemente la discriminación entre apneas e hipopneas. Del mismo modo, Haidar et al. (2018) concluyeron que la detección de eventos respiratorios mejoraba al incluir la presión nasal.

2.1.3 PREPROCESADO DE LA SEÑAL

Cabe destacar también que, en algunos trabajos, en lugar de trabajar directamente con los datos en crudo (*raw data*), se extraen previamente características de la señal antes de ser utilizadas como entradas al modelo. En el caso del ECG, las características extraídas más habituales son los intervalos RR y las amplitudes del complejo QRS (Almutairi et al., 2021a; Sharan et al., 2020; T. Wang et al., 2019; X. Wang et al., 2020). Además, en el estudio de Nasifoglu et al. (2021) emplean transformaciones tiempo-frecuencia como la transformada wavelet y la STFT para obtener los escalogramas y espectrogramas, respectivamente, a partir de dicha señal, y así tratar con datos bidimensionales.

2.1.4 BASES DE DATOS

En cuanto a las bases de datos, la base de datos '**Apnea-ECG**' es una las más recurrentes entre los investigadores en el estudio de la apnea del sueño (Almutairi et al., 2021a; Chang et al., 2020; Dey et al., 2018; Hu et al., 2022; Liu et al., 2023; Mostafa et al., 2020; Mukherjee et al., 2021; Nasifoglu & Erogul, 2021; Qin & Liu, 2022; Sharan et al., 2020; T. Wang et al., 2019; X. Wang et al., 2020; J. Zhang et al., 2021), ya que se encuentra disponible de forma gratuita en https://physionet.org/content/apnea-ecg/1.0.0/.

Esta base de datos contiene un total de 70 sujetos, los cuales fueron divididos equitativamente para el entrenamiento y test de los modelos. Estos registros incluyen la señal de electrocardiograma (ECG) para evaluar la función cardiaca, un conjunto de interpretaciones de apnea realizadas por expertos humanos, y un conjunto de interpretaciones para identificar los complejos QRS. Además, 8 registros cuentan con señales adicionales entre las que se encuentran los movimientos torácicos y abdominales, FA y la SpO_2 (Penzel et al., 2000).

Además del conjunto de datos mencionado anteriormente, existen otras bases de datos que también son utilizadas en los artículos consultados, aunque en menor medida. Entre

ellas se encuentran MESA (Choi et al., 2018a; Haidar et al., 2018) y UCD (Cen et al., 2018; Mostafa et al., 2020; T. Wang et al., 2019).

La base de datos de **UCD** (*College Dublin Sleep Apnea Databse*) contiene 25 registros de polisomnografía en adultos con trastorno de apnea del sueño, de los cuales 21 son varones y 4 son mujeres. Entre las señales disponibles se encuentra el electrocardiograma (ECG), electromiograma (EMG), electrooculografía (EOG), electroencefalografía, los movimientos torácicos y abdominales, la saturación de oxígeno (SpO_2) y registros de ronquidos. Además, cabe destacar que cuenta con una clasificación detallada de las diferentes fases del sueño (Goldberger et al., 2000).

MESA (*Multi-Ethnic Study of Atherosclerosis*) es un estudio de investigación de factores relevantes para el desarrollo de enfermedades cardiovasculares subclínicas en poblaciones multiétnicas y su evolución hacia la etapa clínica en 6814 individuos. Adicionalmente, llevaron cabo pruebas específicas para evaluar el sueño en 2237 participantes, recopilando señales de PSG, datos de actigrafía de muñeca y cuestionarios relacionados con el sueño (Chen et al., 2015; G. Q. Zhang et al., 2018).

Por otro lado, varios investigadores optan por elaborar sus propias bases de datos para el desarrollo de su investigación, en lugar de utilizar conjuntos de datos ya disponibles (Choi et al., 2018a; Erdenebayar et al., 2019; Huttunen et al., 2023; Leino et al., 2021; Nikkonen et al., 2021; Pathinarupothi et al., 2017; Qin & Liu, 2022; Urtnasan et al., 2018). Además, se han visto algunos casos en los que los investigadores emplean varias bases de datos para entrenar y validar sus modelos (Choi et al., 2018a; Leino et al., 2021; Mostafa et al., 2020; Nasifoglu & Erogul, 2021; Qin & Liu, 2022; T. Wang et al., 2019).

Cabe destacar que todas las bases de datos están constituidas por registros de sujetos adultos, siendo la edad mínima de 18 años y la edad máxima de 87 años entre las bases analizadas.

2.1.5 TAMAÑO DE SEGMENTO

Por otra parte, se han podido observar diferentes estrategias en cuando al tamaño de entrada utilizado para la detección de eventos de apnea e hipopnea. Además, varios investigadores han experimentado con diferentes tamaños de segmento para evaluar su impacto en el rendimiento del modelo.

Algunos estudios emplean segmentos de entrada de 10 segundos, lo que proporciona una resolución temporal adecuada para analizar las señales de manera más detallada (Erdenebayar et al., 2019; Urtnasan et al., 2018). Otros optan por tamaños de entrada más largos, como segmentos de 30 y 60 segundos sin solapamiento (Haidar et al., 2018; Nasifoglu & Erogul, 2021; Almutairi et al., 2021a; Chang et al., 2020; Dey et al., 2018; Pathinarupothi et al., 2017; Qin & Liu, 2022; Sharan et al., 2020; T. Wang et al., 2019), lo cual permite capturar información más amplia sobre los eventos respiratorios. Sin embargo, Ling Cen et al. (2018) y Huttunen et al. (2023) se centran en utilizar segmentos de entrada más pequeños, de tan solo 1 segundo, mientras que en el extremo opuesto se encuentra la investigación de Leino et al. (2021) con entradas de 10 minutos.

En algunos estudios también se han probado diferentes tamaños de entrada para determinar cuál ofrece mejores resultados. Por ejemplo, Mostafa et al. (2020) compararon tamaños de 1, 3 y 5 minutos, concluyendo que los segmentos de 3 y 5 minutos mostraron mejores resultados en comparación con los de 1 minuto. Liu et al. (2023), por su parte, evaluaron segmentos de 1 y 3 minutos y observaron un mejor rendimiento con los segmentos de 3 minutos. En consecuencia, concluyeron su trabajo sugiriendo que una ventana de segmentación temporal más larga puede ser beneficiosa para detectar la AOS.

Asimismo, se han encontrado varios enfoques con solapamiento, como en los trabajos de Choi et al. (2018a) y Zhang et al. (2021), donde utilizan segmentos de 10 segundos con solapamiento cada 1 segundo, o en los trabajos de Leino et al. (2021) y Nikkonen et al. (Nikkonen et al., 2021), donde utilizan ventanas de 10 minutos con 98% de solapamiento y ventanas de 30 segundos con 28 segundos de solapamiento, respectivamente.

En términos generales, cuanto mayor es el segmento considerado, más sencillo es detectar un evento de apnea o hipopnea. Sin embargo, a medida que aumenta el tamaño de segmento, también se incrementa la posibilidad de que ocurran varios eventos de apnea o hipopnea dentro de un mismo segmento. Esta situación puede dificultar la posterior estimación del AHI.

2.1.6 DISTINCIÓN ENTRE APNEAS E HIPOPNEAS

En cuanto a la metodología para clasificar los eventos respiratorios por segmentos, se puede observar que la mayoría de los artículos optan por realizar una clasificación binaria entre las clases ‘Normal (N)’ y ‘Apnea (A)’. Sin embargo, con esta aproximación no se distingue entre eventos de apnea e hipopnea, agrupándolos todos dentro de la clase ‘Apnea (A)’. Únicamente en 6 de los artículos consultados llevan a cabo un enfoque de clasificación multiclase entre eventos normales, apneas e hipopneas, incluidos en las clases ‘Normal (N)’, ‘Apnea (A)’ e ‘Hipopnea (H)’ respectivamente (Cen et al., 2018; Erdenebayar et al., 2019; Haidar et al., 2018; Huttunen et al., 2023; Nikkonen et al., 2021; Urtnasan et al., 2018). Entre ellos, Cen et al. (2018) y Nikkonen et al. (2021) y Huttunen et al. (2023) incluyen la señal de SpO_2 junto con otras señales respiratorias.

Cabe destacar que en varios de los artículos, como en el de Liu et al. (2023), mencionan el problema del desequilibrio de clases en los datos de entrenamiento de los modelos. Este desequilibrio se refiere a la situación en la que una de las clases es mayoritaria en comparación con otras, lo que resulta en un número reducido de muestras para las clases minoritarias (generalmente las apneas e hipopneas). Esto podría afectar al rendimiento del modelo sesgando las métricas de evaluación. Para combatir este problema, Qin et al. (2022) llevan a cabo una técnica de sobremuestreo denominada ADASYN, que sintetiza muestras artificiales de las clases minoritarias para mantener los datos balanceados.

2.1.7 ESTIMACIÓN DEL AHI

A parte de la detección por segmento, varios de los artículos analizados se centran en la clasificación por sujeto para determinar la severidad de la AOS (Chang et al., 2020; Choi et al., 2018a; Huttunen et al., 2023; Liu et al., 2023; Mukherjee et al., 2021; Nasifoglu & Erogul, 2021; Nikkonen et al., 2021; Qin & Liu, 2022; T. Wang et al., 2019). Para ello, estiman el AHI utilizando la siguiente ecuación ('Ecuación 2.1'), que tiene en cuenta el número de eventos respiratorios detectados en cada segmento temporal (N) y el número de señales de 60 segundos en un registro completo (L):

$$AHI = \frac{60}{L} x\ N \tag{2.1}$$

2.2 COMPARACIÓN Y ELECCIÓN DEL MÉTODO A IMPLEMENTAR

En la primera parte de este segundo capítulo, se ha llevado a cabo una revisión del estado del arte en la detección automática de eventos de apnea e hipopnea en la AOS utilizando diferentes enfoques de *deep learning* y empleando señales cardiorrespiratorias. Cada uno de estos estudios ha demostrado una gran eficacia en diferentes conjuntos de datos, logrando buenos resultados en la detección de eventos y en algunos casos, en la clasificación por sujeto. Sin embargo, para seleccionar el mejor método a implementar en este TFG, hay que considerar factores como la facilidad de implementación, los recursos computacionales necesarios, el tamaño y la calidad de la base de datos disponible.

En mi caso, se ha decidido implementar una CNN ya que, aunque existen varias opciones, la revisión bibliográfica no muestra diferencias significativas en el rendimiento en comparación con las demás arquitecturas. En consecuencia, se opta por la arquitectura más simple de DL. Asimismo, se va a realizar una clasificación multiclase entre eventos normales, apneas e hipopneas utilizando las señales de SpO_2 y PR de la base de datos de MESA. Esta aproximación es innovadora ya que, hasta la fecha, no se han realizado estudios previos que realicen una clasificación multiclase utilizando dichas señales.

La elección de la base de datos MESA se debe a su elevado número de registros, lo que permitirá obtener resultados más sólidos. Además, se van a utilizar dos tamaños de entrada diferentes: 30 segundos y 60 segundos, ya que son los más utilizados por los autores en los estudios revisados. Esta metodología no ha sido explorada en ninguno de los artículos consultados, y se considera que podría aportar un enfoque novedoso en la detección de eventos respiratorios en la AOS. Adicionalmente, se va a estimar el AHI para poder llevar a cabo una clasificación por sujeto y así determinar la severidad de la AOS en cada paciente.

Por último, como otra contribución con respecto al estado del arte, se va a incluir la aplicación de técnicas de *Explainable Artificial Intelligence (XAI)* para tratar de explicar las decisiones tomadas por el clasificador y aumentar la comprensión del modelo.

CAPÍTULO 3: SUJETOS Y SEÑALES

En este tercer capítulo se presenta la población bajo estudio, la cual está representada por los participantes del "*Multi-Ethnic Study of Atherosclerosis*" (MESA). Además, se profundiza en las características de las señales de pulsioximetría y se describe la metodología utilizada para la división de los datos en entrenamiento, validación y test.

3.1 POBLACIÓN BAJO ESTUDIO: *Multi-Ethnic Study Of Atherosclerosis* (MESA)

El estudio de MESA se diseñó con el objetivo de investigar la prevalencia y progresión de enfermedades cardiovasculares (ECV) subclínicas y la identificación de factores de riesgo de las ECV en una muestra de diferentes grupos raciales y étnicos. Para ello, se reclutaron un total de 6814 sujetos de edades comprendidas entre los 45 y 84 años, que se identificaron como blancos, negro/afroamericanos, hispanos o chinos, y sin ECV aparentes (Chen et al., 2015).

Diez años después se invitó a un determinado grupo de participantes que cumplían ciertos criterios establecidos a participar en un estudio auxiliar denominado 'MESA *Sleep Ancillary Study*'. Concretamente, se excluyeron aquellos pacientes que informaron del uso regular de dispositivos orales, oxígeno nocturno o dispositivos de presión positiva en las vías respiratorias (PAP). Finalmente, hubo un total de 2230 de sujetos que participaron en este nuevo estudio, el cual incluyó la realización de una PSG, una actigrafía y datos de cuestionarios sobre el sueño (Chen et al., 2015). Entre estos participantes, se registraron las señales de pulsioximetría (SpO_2 y PR) en 2056 sujetos.

A partir de la PSG, se anotaron las fases del sueño y los eventos de apnea e hipopnea para posteriormente calcular el AHI (Chen et al., 2015). Los resultados de la PSG revelaron que el 33.8% de los participantes presentaba un trastorno respiratorio del sueño (TRS) de moderado a grave, con un AHI mayor o igual a 15, mientras que el 15% tenía un TRS grave, con un AHI igual o superior a 30. La Tabla 3.1 recoge las características sociodemográficas y clínicas de los participantes del estudio.

	TODOS	**NO AOS**	**AOS LEVE**	**AOS MODERDO**	**AOS GRAVE**
Registros de pulsioximetría (%)	2056 (100%)	215 (10.46%)	622 (30.25%)	609 (29.62%)	610 (29.67%)
Edad promedio (años)	69.3±9	66.6±8.8	68.7±9.0	70.3±8.9	70.0±9.0
Hombres (%)	954	62 (28.83%)	213 (34.24%)	293 (48.11)	386 (63.27)
AHI promedio (e/h)	24.1±19.5	2.7±1.3	9.8±2.9	21.4±4.3	49.0±16.2

Tabla 3.1. Características sociodemográficas y clínicas de los participantes del estudio del sueño de MESA. AOS = apnea obstructiva del sueño.

3.2 CARACTERÍSTICAS DE LAS SEÑALES DE PULSIOXIMETRÍA

Las señales de pulsioximetría (SpO_2 y PR) fueron adquiridas durante la PSG con un sensor de pulsioximetría a una frecuencia de muestreo de 1 Hz. La primera tarea consistió en realizar el preprocesamiento de las señales y las anotaciones relacionadas con el sueño. Esto implica la carga de datos y la división de las señales en segmentos de una duración especificada (30 y 60 segundos para este trabajo). Dado que la saturación de oxígeno en sangre no es una señal de naturaleza eléctrica, no está sujeta a las fuentes de ruido habituales de las señales biomédicas (Cohen, 2006). Sin embargo, en este caso, la principal fuente de ruido es la pérdida de señal por falta de contacto del sensor a causa de movimientos del paciente (Netzer et al., 2001).

Una vez hecho esto, se utilizó el software de MATLAB para visualizar las señales junto con las anotaciones de los eventos respiratorios para así poder detectar posibles errores/inconsistencias en el preprocesado. Concretamente, se trabajó inicialmente con el registro '*mesa-sleep-0033*', por lo que todas las figuras adquiridas se corresponden con dicho registro.

En primer lugar, se representa las señales de SpO_2 y PR, excluyendo las marcas de los eventos, obteniendo así la imagen de la Figura 3.1. No obstante, dada la extensa duración del registro resulta difícil distinguir patrones o características de manera visual. En consecuencia, para facilitar la interpretación se representa también un intervalo de 5 minutos del registro completo, como se muestra en la Figura 3.2. Asimismo, en esta figura se superponen las marcas en aquellos instantes de tiempo en los cuales se produce un evento de apnea o hipopnea.

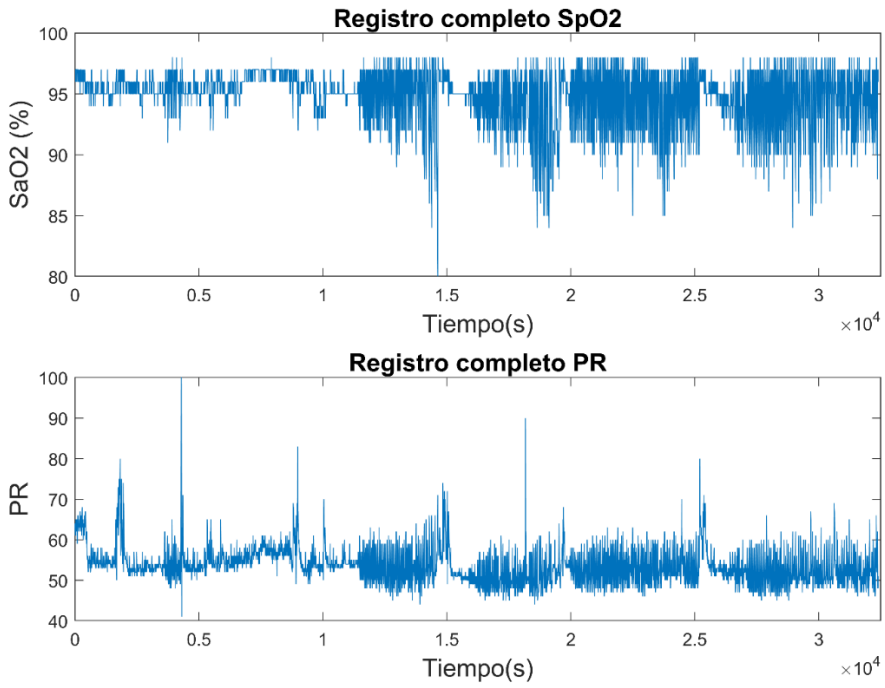

Figura 3.1. Señal de SpO_2 para el registro completo (imagen superior). Señal de pulse rate (PR) para el registro completo (imagen inferior).

En este intervalo se puede distinguir un patrón que consiste en agrupaciones de desaturaciones. Dichos agrupamientos, conocidos como *clusters* de desaturación, se forman cuando varios eventos de desaturación se producen próximos en el tiempo (Vaquerizo-Villar et al., 2021). Además, se puede observar que cuando finaliza el evento de apnea o hipopnea, indicado por las marcas, se produce inmediatamente después una desaturación.

Por otra parte, analizando la señal de pulso se observa que durante las desaturaciones la frecuencia cardiaca se incrementa, lo cual evidencia un intento del organismo de compensar la falta de oxígeno. A su vez, una vez que se inicia la resaturación, la frecuencia de pulso disminuye.

Por último, en la etapa de preprocesado se realizó el etiquetado de los segmentos de la señal a partir de los datos de los eventos de MESA. Para ello se siguió un criterio específico. En los segmentos donde solo había un evento presente se elegía la etiqueta correspondiente a dicho evento, mientras que en los segmentos donde había dos o más eventos no se distinguieron individualmente. Por ejemplo, si se presentaba una apnea y una hipopnea dentro del mismo segmento, se tomaba como etiqueta la del evento que ocurriera primero en el tiempo.

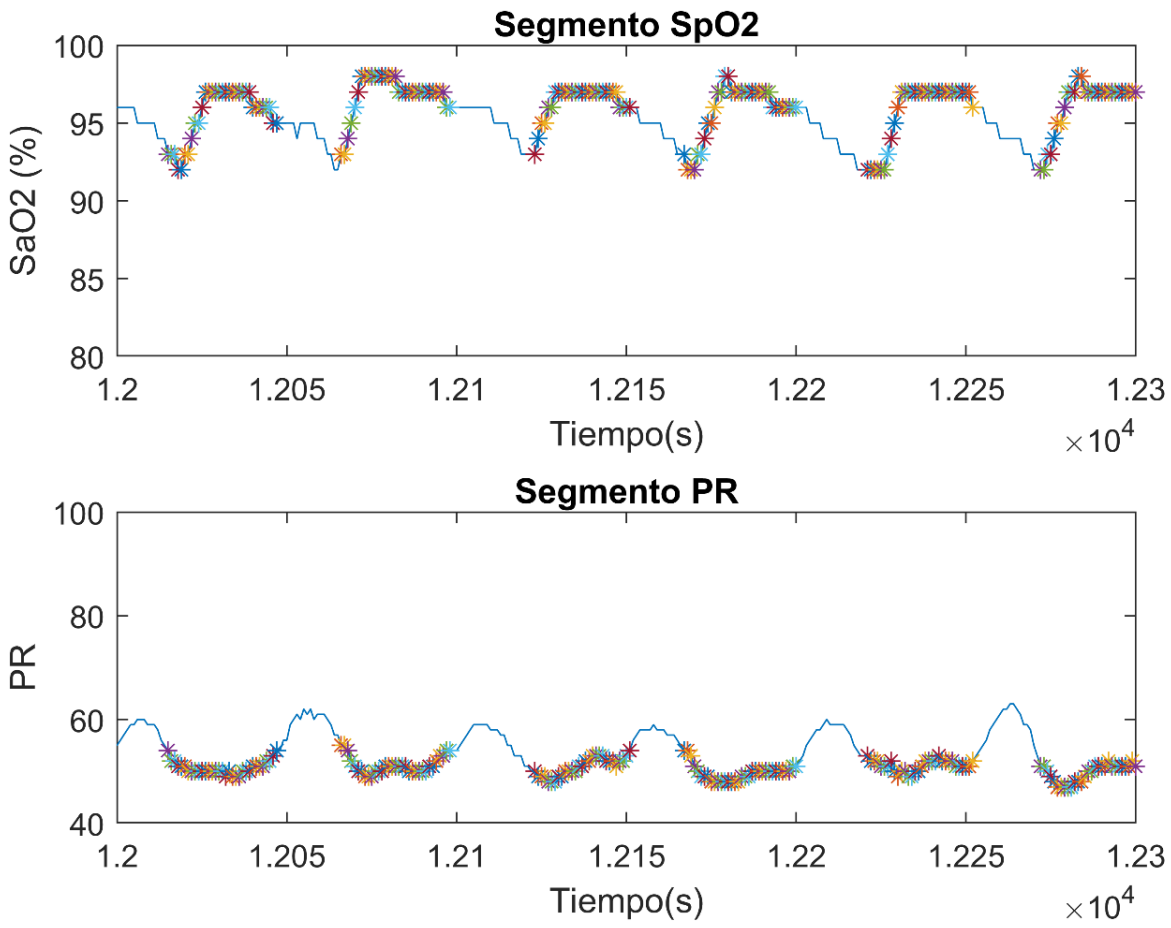

Figura 3.2. Intervalo de 5 minutos para la señal de SpO_2 (imagen superior). Intervalo de 5 minutos para la señal de frecuencia de frecuencia de pulso (imagen inferior).

3.3 DIVISIÓN EN ENTRENAMIENTO, VALIDACIÓN Y TEST

Para diseñar y evaluar modelos de *deep learning* es importante dividir las bases de datos en tres conjuntos diferentes: entrenamiento, validación y test. Además, esta división se realiza de tal forma que cada sujeto esté presente en solo uno de los grupos, garantizando que los datos sean independientes en cada uno de ellos.

El conjunto de entrenamiento es importante para obtener diferentes modelos de clasificación. A medida que avanza el proceso de entrenamiento, los modelos aprenden las características más representativas de los registros que lo componen, lo cual permite al modelo realizar predicciones más precisas. Por otro lado, el conjunto de validación es útil para supervisar, en un grupo independiente, la convergencia y el rendimiento de los modelos durante su entrenamiento. Además, ayuda al programador en la optimización de los hiperparámetros de la red neuronal.

Una vez que los modelos han sido entrenados y evaluados utilizando el conjunto de validación, se selecciona el que demuestra el mejor rendimiento para evaluarlo con el conjunto de test. Esto permite conocer la capacidad de generalización del modelo entrenado y su efectividad en la tarea de clasificación ante nuevos datos de entrada.

Para el desarrollo de este TFG, se dividen los sujetos de la siguiente manera: el 40% de ellas se asignan al conjunto de entrenamiento, mientras que el 30% se destinan a los conjuntos de validación y test, respectivamente. Considerando que hay un total de 2056 registros disponibles en la base de datos MESA tras la etapa de preprocesado, el grupo de entrenamiento está compuesto por 822 registros, mientras que los grupos de validación y test cuentan con 617 cada uno.

Finalmente, se lleva a cabo una etapa de normalización de las señales de SpO_2 y PR por separado. En primer lugar, los datos de entrenamiento se normalizan individualmente calculando la media y la desviación típica de cada conjunto (por cada señal) con el fin de ajustar la escala de valores de cada conjunto en función de su propia distribución. Por otro lado, se emplea la media y desviación típica previamente calculadas en los datos de entrenamiento para normalizar los datos de los segmentos de validación y test.

CAPÍTULO 4: METODOLOGÍA

4.1 REDES NEURONALES CONVOLUCIONALES

En el campo del *deep learning*, las redes neuronales convolucionales destacan como uno de los algoritmos más reconocidos y ampliamente utilizados. Esto es debido principalmente a su habilidad para identificar de manera automática las características relevantes en los datos sin necesidad de requerir intervención humana (Ye, 2022). En consecuencia, las aplicaciones de las CNN se han extendido de forma considerable en los últimos años, incluyendo el análisis y clasificación de series temporales (Ismail Fawaz et al., 2019).

4.1.1. ESTRUCTURA GENERAL

Al igual que las ANN convencionales, la arquitectura de la CNN se inspira en las conexiones neuronales presentes en el cerebro humano. Esta estructura típicamente está compuesta por múltiples capas convolucionales seguidas por capas de submuestreo o *pooling*, y finalmente, capas *fully-connected* (FC) (Saxena, 2022). En la Figura 4.1 se representa un ejemplo de CNN simplificada para clasificar imágenes de la base de datos MNIST.

La entrada x de cada capa tiene una estructura tridimensional: altura (*height*), ancho (*width*) y profundidad (*depth*), denotadas como $h\ x\ w\ x\ d$, donde la profundidad d representa el número de canales. Además, en cada capa convolucional se utilizan múltiples filtros o *kernels* representados por k, los cuales tienen cada uno un *bias* b^k y unos pesos W^k determinados para generar k mapas de características h^k. A su vez, los filtros tiene también tres dimensiones ($f\ x\ f\ x\ d$) (Alzubaidi et al., 2021). Es importante destacar que la profundidad del filtro 'd' es siempre la misma que la de la entrada de la capa a la que se aplica (Ye, 2022).

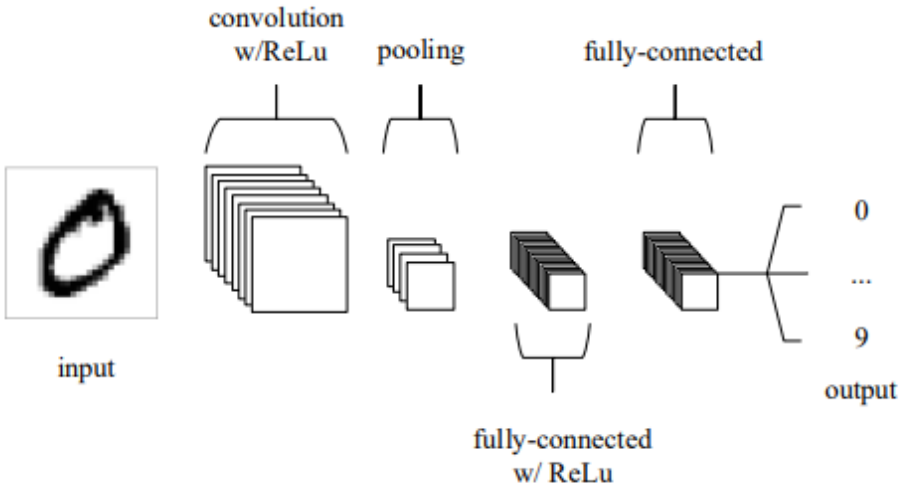

Figura 4.1. Arquitectura CNN compuesta por una capa de entrada, una capa convolucional, una capa pooling, dos capas fully-connected y una capa de salida (Saxena, 2022).

4.1.1.1. CAPAS CONVOLUCIONALES

En la estructura de una CNN, uno de los elementos destacados son las capas convolucionales. Estas capas están compuestas por una serie de filtros de convolución, también denominados *kernels* de convolución, conteniendo cada *kernel* k un bias (b^k) y de la matriz de pesos (W^k), que desempeñan un papel fundamental en la generación del mapa de características h^k a partir de la entrada. Al inicio del proceso de entrenamiento de la CNN se les asignan valores aleatorios que posteriormente son ajustados en cada una de las épocas de entrenamiento. En consecuencia, el núcleo aprende a extraer características significativas durante el proceso de aprendizaje (Alzubaidi et al., 2021).

En las capas convolucionales se realiza una operación de convolución, que consiste en colocar el filtro en cada posición posible de la entrada o capa oculta, deslizándolo a lo largo de todas las direcciones. Luego, se calcula el producto elemento a elemento entre los parámetros del filtro y la cuadrícula correspondiente, multiplicando sus valores y sumándolos para obtener un único valor escalar (Ye, 2022). Estos valores calculados representan el mapa de características de salida (Alzubaidi et al., 2021). En la Figura 4.2 se ilustran gráficamente los cálculos realizados, donde el *kernel* y el área correspondiente de la imagen de entrada se multiplican, y el resultado se suma para generar el mapa de características de salida (Kim, 2017, Chapter 6).

4.1.1.2. CAPAS POOLING

A continuación, se reduce el tamaño de los mapas de características creando versiones más compactas mediante las capas *pooling*, lo cual acelera el proceso de entrenamiento y previene el *overfitting*. A pesar de esta reducción, se busca preservar las características más relevantes de los datos. Para lograr esto se pueden aplicar diferentes métodos como el *max pooling, min pooling*, el *average pooling* o el *global average pooling*, entre otros, en un área adyacente de tamaño pxp, donde p es el tamaño del kernel (Alzubaidi et al., 2021). Los más comunes y ampliamente utilizados son el *max pooling* y el *average pooling*: el *max pooling* toma el valor máximo mientras que el *average pooling* calcula el promedio de los valores en el área correspondiente, como se representa en la Figura 4.3, donde se hace *pooling* por cuadrículas de 2x2 (Kim, 2017, Chapter 6).

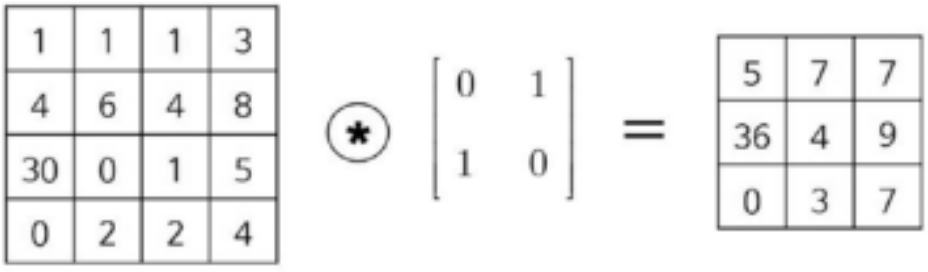

Figura 4.2. Ejemplo de operación de convolución (Kim, 2017, Chapter 6).

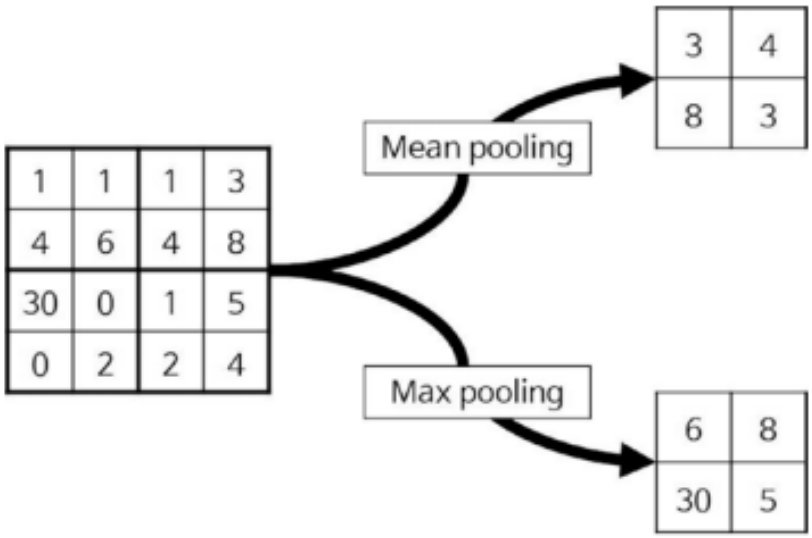

Figura 4.3. Ejemplo de max pooling y average/mean pooling (Kim, 2017, Chapter 6).

4.1.1.3. FUNCIONES DE ACTIVACIÓN

En todas las redes neuronales la función de activación f desempeña un papel muy importante al mapear la entrada a la salida. Al procesar una neurona, se calcula la suma ponderada de la entrada ($W^k * x$) y su bias b^k. Después, la función de activación entra en acción tomando la decisión de activar o desactivar la neurona para generar la salida correspondiente. Por tanto, al aplicar una función de activación se obtiene un mapa de características con la siguiente expresión (Ecuación 3.1) (Alzubaidi et al., 2021):

$$h^k = f(W^k * x + b^k) \tag{4.1}$$

En la arquitectura de las CNN se utilizan funciones de activación no lineales después de todas las capas con pesos, como las capas *fully-connected* y las capas convolucionales. Esto implica que el mapeo entre la entrada y la salida será no lineal, otorgando a la CNN la capacidad de aprender patrones más complejos. Además, las funciones de activación deber ser diferenciables para permitir el uso del algoritmo de *backpropagation* durante el proceso de entrenamiento de la red y así optimizar el rendimiento de la CNN (Alzubaidi et al., 2021).

Entre los tipos más comunes de funciones de activación utilizadas en las CNN se encuentran: sigmoide, tangente hiperbólica (tanh), ReLU (*Rectified Linear Unit*), *Leaky* ReLU, PReLU y ELU (Rasamoelina et al., 2020). Concretamente, una de las funciones más frecuentemente empleadas es la ReLU, cuya expresión matemática se describe en la Ecuación 4.2 (Alzubaidi et al., 2021).

$$f(x)\ ReLU = max\ (0, x) \tag{4.2}$$

Esta función elimina las entradas negativas y mantiene las entradas positivas sin realizar ninguna modificación, lo que le otorga su no linealidad. Además, una de sus principales ventajas en comparación con otras funciones de activación es su eficiencia computacional, lo que favorece su amplia variedad de uso en aplicaciones de *deep learning* (Kim, 2017, Chapter 5; Ye, 2022).

4.1.1.4. CAPAS FULLY-CONNECTED

Las capas *fully-connected* (FC) comúnmente se ubican al final de cada arquitectura CNN. En esta capa cada neurona está conectada con todas las neuronas de la capa anterior, siguiendo el modelo del perceptrón multicapa, que es un tipo de ANN *feed-forward*. La entrada de la capa FC recibe las características de bajo nivel de la última capa convolucional o capa *pooling*. Finalmente, la salida de la capa FC representa la salida final de la CNN (Alzubaidi et al., 2021).

Por lo general, se utilizan capas FC con una función de activación ReLU en las capas intermedias de una red neuronal, con el fin de introducir no linealidades y capturar características importantes del conjunto de datos. Asimismo, en tareas de clasificación multiclase, se opta por emplear la función *softmax* en la capa de salida. La razón radica en que esta función permite asignar una probabilidad a cada clase objetivo en un rango de 0 a 1. Para ello, toma el vector de salida de la última capa FC y lo transforma en una distribución de probabilidad, donde cada valor representa la probabilidad de pertenecer a una clase específica. Esta característica permite elegir la clase con la mayor probabilidad como la predicción final del modelo. (Kim, 2017, Chapter 4).

4.1.2. VENTAJAS SOBRE ANN CONVENCIONALES

Los estudios de Lecun et al. (2015) destacan tres características fundamentales de las CNN que mejoran significativamente el rendimiento respecto a las ANN convencionales: representaciones equivalentes (*equivariant representations*), interacciones dispersas (*sparse interactions*) y el uso compartido de parámetros (*parameter sharing*).

En las ANN convencionales, cada neurona de una capa está conectada con todas las neuronas de la siguiente capa. Sin embargo, en las CNN se establecen pocas conexiones entre capas adyacentes, asignándose solo unos pocos pesos, de manera que la memoria requerida para almacenar los pesos es menor. Esta característica, conocida como ***sparse interactions***, hace que las CNN sean más eficientes en términos de memoria y coste computacional (Alzubaidi et al., 2021).

Como se ha comentado, en las CNN se utiliza el concepto de '***parameter sharing***', que se refiere al uso del mismo parámetro para más de una función en un modelo (Lecun et al., 2015). Esto consiste en operar con un conjunto de pesos para todos los píxeles de la entrada en vez de asignar pesos individuales, como ocurre en las ANN (Alzubaidi et al., 2021). Por lo tanto, en lugar de aprender un conjunto independiente de parámetros para cada píxel, solo se aprende un conjunto único. Esta estrategia permite disminuir significativamente el tiempo de entrenamiento y agilizar el proceso de aprendizaje de la CNN respecto a las ANNs (Lecun et al., 2015). Asimismo, el hecho de compartir parámetros otorga a las CNN la característica de invarianza a la traslación '***equivariant representations***'. Esto implica que, si la entrada experimenta algún cambio, la salida también será modificada de manera equivalente (Lecun et al., 2015). Todo ello, junto con la capacidad de las CNN para trabajar directamente con los datos en bruto (*raw data*), sin requerir una selección manual previa de características, hace que tengan un mejor rendimiento en tareas de procesamiento de imágenes y señales temporales en comparación con ANN tradicionales (Lecun et al., 2015).

4.2 ARQUITECTURA CNN APLICADA

Como se vio en el Capítulo 2, los eventos respiratorios que se producen en la AOS se pueden clasificar en dos categorías: N/A (normal-apnea), incluyéndose dentro del grupo 'apnea' tanto las apneas como las hipopneas. Alternativamente, también se puede realizar una clasificación multiclase discriminando entre apneas (A), hipopneas (H) y respiración normal (N), utilizando la nomenclatura N/A/H, que es la elección para este TFG. Para llevar a cabo esta tarea de clasificación multiclase se ha implementado una CNN en Python utilizando la librería Te*nsorFlow*, una biblioteca de código abierto para crear modelos de *machine learning*. A continuación, se van a detallar las principales características del modelo comenzando por las entradas a la red neuronal, siguiendo con su arquitectura y terminando con las técnicas de regularización y optimización empleadas.

4.2.1 ENTRADAS A LA RED

Para este trabajo, las entradas de la red neuronal están constituidas por los segmentos de 30/60 segundos de las señales de pulsioximetría (SpO_2 y PR) extraídas de la base de datos MESA. Estos tamaños de segmento son los más utilizados por la comunidad científica en la detección de eventos de apnea e hipopnea, demostrando ser efectivos para capturar información relevante (Almutairi et al., 2021b; Chang et al., 2020; Choi et al., 2018b; Haidar et al., 2018; Mostafa et al., 2020; Mukherjee et al., 2021; Nasifoglu & Erogul, 2021; Nikkonen et al., 2021; Pathinarupothi et al., 2017; Qin & Liu, 2022; Sharan et al., 2020; X. Wang et al., 2020).

Para cada tamaño de segmento, se exploran diferentes tipos de señales de entrada para poder comparar el rendimiento del modelo en cada caso. En primer lugar, se utiliza un solo canal de entrada con la señal de SpO_2. Luego, se prueba un único canal con la señal de PR. Posteriormente, se combinan ambas señales, con el fin de evaluar si al utilizar las dos señales de pulsioximetría se logra mejorar la detección de eventos de apnea e hipopnea.

Además, se explora el rendimiento del modelo al introducir adyacencia de 1 y 2 segmentos. Como se ilustra en la Figura 4.4, en el caso de la adyacencia de 1 segmento se considera el segmento correspondiente, el anterior y el posterior, utilizando la etiqueta del segmento central. De forma similar, en la adyacencia de 2 segmentos se toman los dos segmentos previos y los dos siguientes, manteniendo la etiqueta del segmento central. Esta estrategia permite que la secuencia temporal de entrada sea más extensa, alcanzando un máximo de 150 segundos para el tamaño de segmento de 30 segundos y un máximo de 300 segundos para los segmentos de 60 segundos. Además, puede ayudar a mitigar el efector negativo del retardo variable existente entre la desaturación y la ocurrencia de la apnea/hipopnea.

Por lo tanto, al realizar pruebas con las configuraciones de entrada comentadas, se pretende analizar cómo afecta la longitud y el tipo de señal en el rendimiento general de la red neuronal para detectar eventos de apnea e hipopnea y realizar una clasificación por sujeto. En la Figura 4.5 se representan de forma esquemática todas las entradas consideradas.

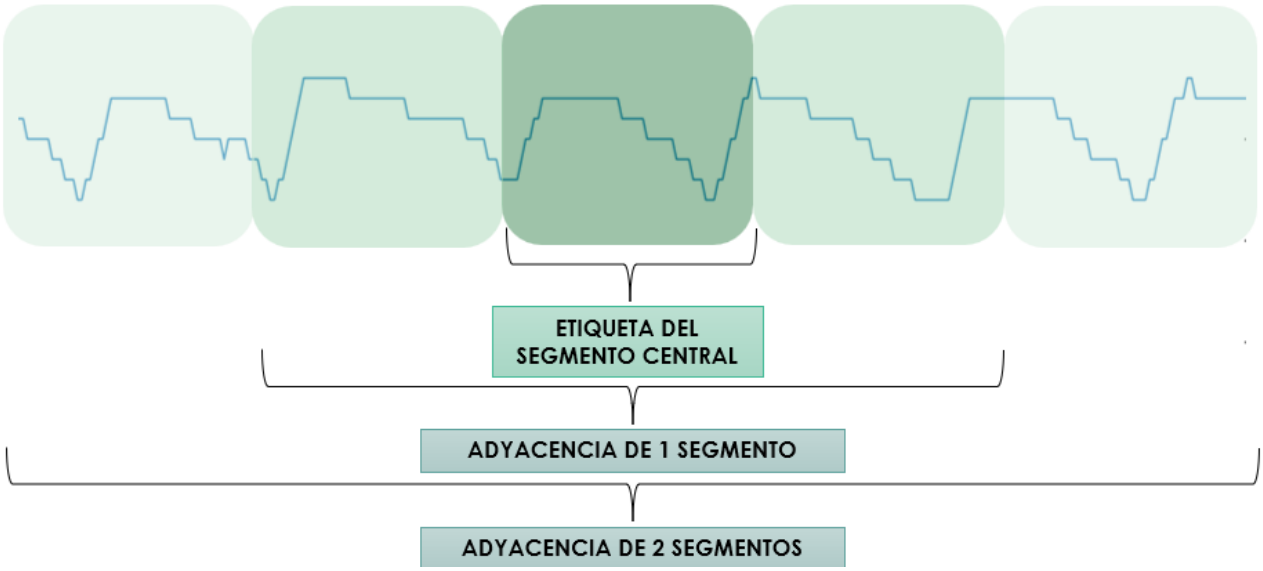

Figura 4.4. Adyacencia.

Finalmente, al analizar los segmentos disponibles para entrenamiento y validación se observó que solo un 17,6% contenía eventos de apnea o hipopnea en ambos conjuntos, mientras que la gran mayoría de las muestras pertenecía a la clase 'normal'. Este desbalanceo de clases podría generar resultados engañosos durante la evaluación del modelo obteniendo una exactitud elevada al clasificar correctamente los eventos normales debido a su mayor representación en los datos. Sin embargo, la detección de apneas e hipopneas podría ser mucho menos precisa debido a la escasez de ejemplos de estas clases en el conjunto de datos.

Para abordarlo, se ha aplicado un balanceo de clases en los conjuntos de entrenamiento y validación. Este proceso se ha llevado a cabo mediante el uso de sobremuestreo u *oversampling* con la función '*oversampler'* de la librería '*RandomOverSampler'*. Con esta técnica se ha conseguido aumentar el número de ejemplos de las clases minoritarias mediante la replicación de muestras existentes hasta alcanzar una proporción del 66,66% entre las clases 'apnea' e 'hipopnea'.

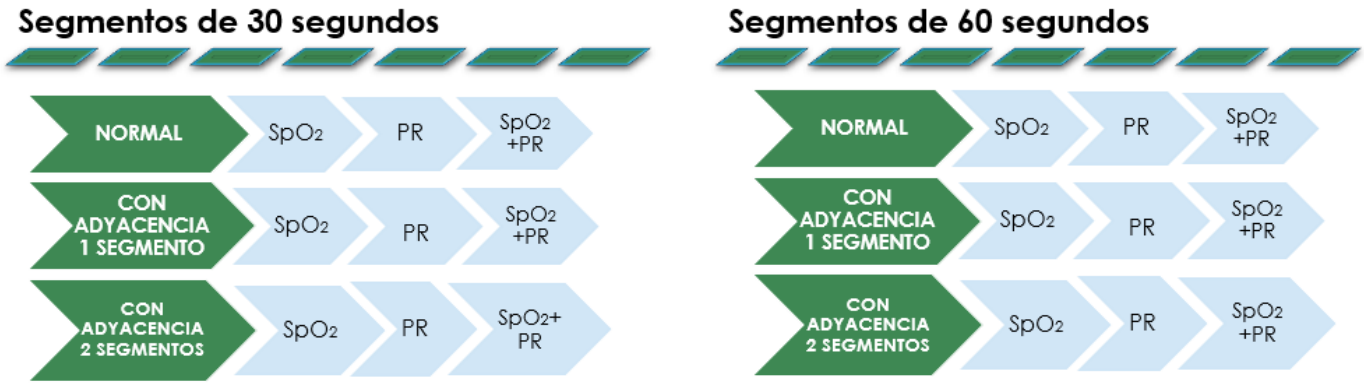

Figura 4.5. Diferentes entradas a la red neuronal.

4.2.2 ARQUITECTURA

Según lo comentado anteriormente, la arquitectura implementada es una CNN para clasificación multiclase de eventos respiratorios en el contexto de la AOS, tal como se representa en la Figura 4.6. Esta red consta de varias capas que se describen a continuación:

Sección de entrada: constituida por una capa de entrada y *Batch-Normalization* para estandarizan los valores de la señal de entrada. La capa de entrada tiene una forma (L, c), donde L denota la longitud del segmento de entrada (30 o 60 segundos) y c se corresponde con el número de canales (1 o 2, dependiendo se si trabaja con las señales de pulsioximetría individualmente o combinadas).

Sección CNN: esta parte de la arquitectura está formada por cuatro capas convolucionales para extraer características relevantes de la señal de entrada, con una función de activación ReLU para introducir no linealidades en el modelo. Asimismo, cada una estas capas convolucionales es seguida por una capa de *batch-normalization*. A continuación, se aplican las capas de *MaxPooling* para reducir la dimensionalidad de las características extraídas y seleccionar únicamente las más importantes. También se emplea una capa de *dropout* con una tasa de 0.1 tratando de evitar que se produzca *overfitting*.

En esta sección de la arquitectura se utilizan filtros de distintos tamaños para realizar la operación de convolución. Concretamente, las dos primeras capas convolucionales utilizan un filtro de tamaño 5, mientras que las dos capas restantes emplean filtros de tamaño 3 con 16 y 32 canales, respectivamente. Además, el parámetro '*padding'* se establece en '*same'* y el parámetro '*stride'* se fija en 1 en todas las capas convolucionales para asegurar que se preserven las dimensiones de la señal a medida que se van realizando las operaciones de convolución.

Sección de salida y clasificación: la última parte de la arquitectura está compuesta por capas densas (*fully-connected*). En primer lugar, se utiliza una capa FC con 32 unidades y función de activación ReLU. Finalmente, se tiene una capa de salida FC con 3 unidades y función de activación *softmax*, que representa las tres clases que se buscan detectar: respiración normal, apnea e hipopnea.

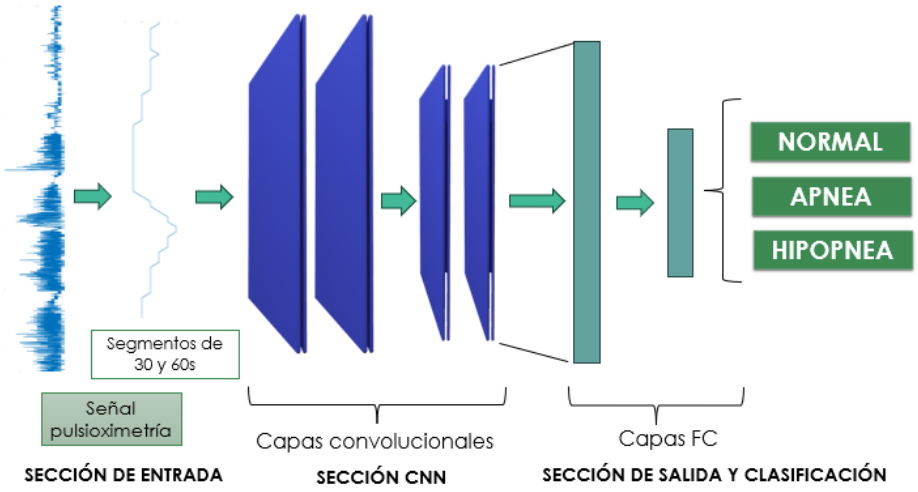

Figura 4.6. Arquitectura de la CNN aplicada para la detección de eventos de apnea e hipopnea.

En el caso de utilizar dos canales de entrada, se han probado dos estrategias distintas: utilizar convoluciones 1D y convoluciones 2D, lo cual implica tener que ajustar la arquitectura. Al emplear **convoluciones 1D,** cada canal se considera por separado y se aplican individualmente las convoluciones, la función de activación, el *batch-normalization* y el *pooling*. Posteriormente se fusionan los mapas de características para continuar con el resto de la arquitectura, como se describió anteriormente. En el caso de **convoluciones 2D**, las señales de SpO_2 y PR se tratan como dos canales de una imagen. En consecuencia, hay que utilizar filtros de dos dimensiones para las convoluciones y capas *MaxPooling2D* para reducir la dimensionalidad. Finalmente, hay que comentar que para ambas estrategias la última sección de la arquitectura permanece idéntica, ya que la información de los canales se procesa de forma similar independientemente de la estrategia utilizada.

4.2.3 REGULARIZACIÓN Y OPTIMIZACIÓN DE LA CNN

En las CNN, el principal problema asociado con la obtención de un modelo generalizable es el sobreajuste u *overfitting*. El *overfitting* ocurre cuando el modelo actúa excepcionalmente bien en los datos de entrenamiento, pero falla en los datos de test, es decir, en un conjunto de datos no vistos previamente. En el extremo opuesto están los modelos que no logran aprender lo suficiente con los datos de entrenamiento, lo que se conoce como *underfitting*. Por lo tanto, el objetivo es implementar un modelo correctamente ajustado que se desempeñe bien tanto en los datos de entrenamiento como en los datos de test (Alzubaidi et al., 2021).

Para evitar el problema del sobreajuste en la CNN implementada, se han aplicado varias técnicas de regularización: *dropout*, *batch normalization*, *early stopping* y *reduceLROnPlateau*.

En primer lugar, *dropout* es una técnica empleada para mejorar la capacidad de generalización del modelo que consiste en desactivar temporalmente neuronas de forma aleatoria en cada época de entrenamiento para evitar que se vuelvan demasiado dependientes entre sí (Srivastava et al., 2014). Concretamente, en este modelo se aplica un *dropout* con una tasa del 0.1, por lo que aproximadamente el 10% de las neuronas se desactivan de forma aleatoria durante el entrenamiento (Vaquerizo-Villar et al., 2021).

El *batch normalization* consiste en normalizar los mapas de características extraídos en cada bloque convolucional, restando la media y dividiendo entre la desviación típica a lo largo de las muestras de un *batch*, lo cual garantiza un procesamiento de la información más uniforme. Además, tiene la ventaja adicional de reducir la variabilidad de la covarianza interna en las capas de activación (Alzubaidi et al., 2021). En la arquitectura propuesta se aplica el *batch normalization*, tanto en la capa de entrada como después de cada capa convolucional, teniendo en cuenta que el tamaño de *batch* elegido es de 256.

La técnica del *early stopping* se emplea para detener el entrenamiento cuando no se observa una mejora significativa en el conjunto de validación durante un determinado número de épocas (Kukačka et al., 2017). En este caso se ha establecido una paciencia de 10 épocas, valor determinado experimentalmente, y se ha configurado el parámetro

'*restore_best_weights*' para que se restauren los mejores pesos obtenidos durante el entrenamiento.

Por último, la técnica *ReduceLROnPlateau* se utiliza para reducir gradualmente la tasa de aprendizaje o *learning rate (LR)* cuando el rendimiento en el conjunto de validación deja de mejorar. Esta función tiene el parámetro '*factor'* para indicar la cantidad que se irá reduciendo la LR, el cual se ha establecido en 0.1; el parámetro '*patience'* que se ha configurado en 5 épocas; y el parámetro '*min_lr'* para establecer el valor mínimo que puede tomar la LR, en este caso 0.000001. Estos valores han sido determinados de manera experimental.

En lo que respecta a la optimización del modelo, en la revisión del estado del arte se observó que la gran mayoría de los modelos empleaban el optimizador *Adaptative Moment Estimation* (Adam) y la función de pérdida '*categorical_crossentropy'*. Por ello, siguiendo su ejemplo, en este modelo se utilizan ambos. Por un lado, Adam es un método de optimización que destaca por su capacidad para calcular tasas de aprendizaje adaptativas para cada parámetro del modelo, permitiendo una convergencia más rápida (Ruder, 2016). Por otro lado, la función de coste permite monitorizar el error entre las predicciones del modelo y las etiquetas reales del conjunto de datos (Ciampiconi et al., 2023).

Por último, dada la variabilidad en el número de capas convolucionales utilizadas por los autores en sus modelos, se ha optimizado este hiperparámetro utilizando el conjunto de validación. Inicialmente se probó con 3 capas convolucionales y se fueron añadiendo capas adicionales hasta llegar a un total de 8. Los resultados mostraron que al aumentar el número de capas no se producían mejoras en el rendimiento, sino que empeoraba. Además, la red se volvía más compleja, incrementándose el número de parámetros y, en consecuencia, se aumentaba considerablemente el tiempo de entrenamiento. Finalmente, el modelo alcanzó los mejores resultados en el conjunto de validación utilizando 4 capas convolucionales.

4.2.4 ESTIMACIÓN DEL AHI

Tras estimar el número de eventos de apnea e hipopnea en segmentos de 30 y 60 segundos, se calculó la tasa de eventos respiratorios. Para ello se sumaron los eventos de apnea e hipopnea detectados y el resultado se dividió entre el tiempo total de registro. Sin embargo, es importante señalar que este cálculo tiene a subestimar el AHI, ya que el tiempo total de registro suele ser mayor que el tiempo real de sueño. Además, hay algunos eventos de apnea que no se asocian con desaturaciones y, en consecuencia, no pueden ser detectados por la CNN (Vaquerizo-Villar et al., 2021).

Para corregir esta subestimación, se empleó un enfoque basado en una regresión lineal, siguiendo la expresión de la Ecuación 4.3, donde β y ε son los coeficientes y la '*disturbance'*, respectivamente. Cabe destacar que el modelo de regresión lineal (los valores de β y ε) fue obtenido utilizando el conjunto de entrenamiento.

$$AHI = (\beta \cdot y_{CNN}) + \varepsilon \tag{4.3}$$

4.3 EXPLAINABLE ARTIFICIAL INTELLIGENCE

En la actualidad, la inteligencia artificial ha adquirido un gran poder en la toma de decisiones en diversos ámbitos de la vida cotidiana: desde recomendaciones personalizadas en plataformas de entretenimiento hasta asistentes virtuales de apoyo en aplicaciones online (Adadi & Berrada, 2018). No obstante, la toma de decisiones en el diagnóstico de enfermedades como la AOS tiene un impacto significativo, por lo que se vuelve esencial conocer las razones detrás de una determinada decisión.

A pesar de que los algoritmos de inteligencia artificial están dotados de grandes capacidades en términos de resultados y predicciones, sufren de opacidad (*black-box perception*), lo cual dificulta la obtención de información acerca de su funcionamiento interno, especialmente en los algoritmos de *deep learning*. Además, confiar decisiones importantes a un sistema que no puede explicarse a sí mismo puede resultar peligroso (Adadi & Berrada, 2018). En consecuencia, para abordar este desafío surge la *explainable articial intelligence* tratando de desarrollar técnicas que permitan generar modelos explicables sin comprometer su rendimiento. Asimismo, buscan asegurar que sean accesibles, transparentes y confiables para su uso en diversos entornos y aplicaciones (Barredo Arrieta et al., 2020). Este concepto de explicabilidad está estrechamente ligado al concepto de interpretabilidad, ya que los sistemas interpretables son explicables cuando sus operaciones pueden ser entendidas por seres humanos (Adadi & Berrada, 2018).

En el contexto de la AOS, ninguno de los artículos del estado del arte incorpora técnicas de XAI, mencionando algunos esta posibilidad como una línea de investigación futura. Por ello, en este TFG sería interesante tratar de comprender las reglas (características de las señales de SpO_2 y PR) que llevan al algoritmo (CNN) a decidir si un determinado evento respiratorio es una apnea o una hipopnea. En consecuencia, se va a implementar la técnica '*Gradient-weighted Class Activation Mapping*' (Grad-CAM) por su efectividad a la hora de visualizar las regiones de la señal que más contribuyen a la decisión de la CNN (Loh et al., 2022). Esto va a permitir identificar los posibles factores que influyen en el rendimiento del modelo.

Grad-CAM es una técnica propuesta para dotar de interpretabilidad a los modelos basados en *deep learning* sin necesidad de modificar su arquitectura, a diferencia de lo que ocurría con la técnica *Class Activation Mapping* (CAM) que obligaba a suprimir las capas FC comprometiendo la exactitud de los modelos. Por lo tanto, Grad-CAM es una generalización del enfoque CAM que puede ser aplicado a una gran variedad de modelos CNN, incluyendo las redes convolucionales con capas FC como es el caso de este TFG (Selvaraju et al., 2020).

En los últimos años, varios estudios han investigado la capacidad de las capas más profundas de las CNN para identificar características visuales de alto nivel (Bengio et al., 2013). Por ejemplo, se ha observado que las capas convolucionales retienen de manera natural información mucho más significativa para el análisis visual en comparación con las *capas fully-connected* (Selvaraju et al., 2020). En consecuencia, para obtener el mapa de calor, Grad-CAM realiza un cálculo que implica el uso de gradientes. Estos gradientes se obtienen al evaluar la salida de la red neuronal y^c en relación con los mapas de características bidimensionales de una capa convolucional específica $A_{i,j}^k$.

Posteriormente, estos gradientes se combinan y promedian para producir una serie de pesos α_k^c. Estos pesos juegan un papel crucial al determinar la importancia relativa de cada mapa de características k en relación con la clase objetivo c, como se expresa en la Ecuación 4.4. En otras palabras, esos pesos α_k^c indican qué regiones o características específicas son más relevantes para la clasificación de la clase en cuestión (Selvaraju et al., 2020).

$$\alpha_k^c = \frac{1}{Z}\sum_i \frac{\partial y^c}{\partial A_i^k} \tag{4.4}$$

Z denota el número de mapas de características presentes en la capa seleccionada. A continuación, se calcula el mapa de calor mediante una operación de ponderación y se combinan estos mapas de características. Además, se aplica una función de activación ReLU a este proceso (Ecuación 4.5) (Selvaraju et al., 2020).

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c \bullet A^k\right) \tag{4.5}$$

En definitiva, este enfoque implica calcular el gradiente mediante *backpropagation* de la puntuación de cada clase (N/A/H en este TFG) con respecto a las activaciones del mapa de características de una capa convolucional. Además, se realiza un promedio global sobre sus dimensiones para obtener los pesos de mayor importancia. Todo ello permite conocer las regiones del mapa de características que son más relevantes para llevar a cabo la clasificación en una determinada clase.

Aunque Grad-CAM puede aplicarse a cualquier capa convolucional de la CNN, en este TFG únicamente nos vamos a enfocar en explicar las decisiones tomadas por la última capa convolucional de la red, que es lo habitual en arquitecturas CNN (Selvaraju et al., 2020). Con esto se pretende ver cómo llega el modelo a sus conclusiones finales y ver qué detalles específicos de las señales de pulsioximetría influyen en cada decisión. Para ello se va a utilizar el modelo que demuestre un buen rendimiento en la clasificación de eventos de apnea e hipopnea. Partiendo de esa premisa, se seleccionarán varios segmentos que hayan sido correctamente clasificados con la probabilidad más alta para cada clase (N/A/H). Después, a partir de estos segmentos se calcularán los *heatmaps* considerando las distintas señales de entrada al modelo (SpO_2, PR y SpO_2+PR) para poder comparar si varían las regiones influyentes en la decisión de cada clase en función de la señal utilizada.

4.4 ANÁLISIS ESTADÍSTICO

Para medir el rendimiento de un modelo en la tarea de clasificación, se emplean una serie de métricas derivadas de la matriz de confusión (*confusion matrix*). La matriz de confusión es una tabla que muestra las clases predichas por el modelo en comparación con las clases reales. Asimismo, esta herramienta puede ser utilizada tanto para clasificación binaria como para clasificación multiclase, permitiendo obtener distintas métricas como la exactitud (*accuracy*, Acc), la sensibilidad (Se) o *recall* y la especificidad (Sp), entre otras (Grandini et al., 2020). Para ello es necesario definir los siguientes elementos de la matriz de confusión (Krstinić et al., 2020):

- Verdaderos positivos (VP): número de casos positivos que el modelo clasifica correctamente como positivos.
- Verdaderos negativos (VN): número de casos negativos que el modelo clasifica correctamente como negativos.
- Falsos positivos (FP): número de casos negativos que el modelo clasifica erróneamente como positivos.
- Falsos negativos (FN): número de casos positivos que el modelo clasifica erróneamente como negativos.

En este TFG se llevan a cabo dos tareas de clasificación multiclase. En la primera de ellas, denominada clasificación por segmento, el modelo tiene que distinguir entre tres clases: 'N' cuando el paciente no presenta eventos respiratorios en ese segmento, 'A' en aquellos segmentos en los que tiene lugar una apnea y 'H' cuando se produce una hipopnea. Por otro lado, en la tarea de clasificación por sujeto, cada individuo es clasificado en función de la severidad de la AOS en cuatro clases: no AOS, AOS leve, AOS moderada y AOS grave, basándose en el AHI. En ambas tareas se van a construir y analizar las matrices de confusión para extraer diferentes estadísticos que permitan evaluar de manera específica el rendimiento del modelo.

4.4.1. MÉTRICAS DE RENDIMIENTO DE LA CLASIFICACIÓN DE EVENTOS DE APNEA E HIPOPNEA

Para poder evaluar el rendimiento de los modelos implementados en la detección de eventos de apnea e hipopnea, se van a calcular dos métricas: el Acc por segmento y el F1-*score*. Por un lado, la **exactitud por segmento** mide la proporción de segmentos correctamente clasificados en relación con el número total de segmentos evaluados, como se define en la Ecuación 4.6 (Grandini et al., 2020). Sin embargo, como se comentó anteriormente, esta métrica puede proporcionar resultados engañosos ante un conjunto de datos desbalanceado.

$$Acc\ por\ segmento = \frac{VP+VN}{VP+VN+FP+FN} \tag{4.6}$$

Por otro lado, el **F1-*score*** es una métrica que combina la precisión (Ecuación 4.7) y el *recall* (Ecuación 4.8), que sirve para evaluar de forma global el rendimiento de un modelo. Concretamente, mide la capacidad del modelo para detectar casos positivos (*recall*) y ser preciso en las clasificaciones que realiza (*precision*). El F1-*score* normalmente se expresa en una escala de 0 a 1, donde 1 representa el mejor desempeño posible. Para calcularlo se utiliza la Ecuación 4.9 (Grandini et al., 2020).

$$precision = \frac{VP}{VP+FP} \tag{4.7}$$

$$recall = \frac{VP}{VP+FN} \tag{4.8}$$

$$F1-score = \frac{2\ precision\ x\ recall}{precision+recall} \tag{4.9}$$

4.4.2. MÉTRICAS DE RENDIMIENTO DE LA ESTIMACIÓN DEL AHI

Para evaluar el rendimiento de los diferentes modelos en la estimación del AHI para clasificar los sujetos en función de la severidad de la AOS, se van a calcular varias métricas: la exactitud, el coeficiente kappa de cohen (kappa, k) y el coeficiente de correlación intraclase (*intraclass correlation coefficient*, ICC).

En primer lugar, la **exactitud por sujeto** permite medir qué proporción de sujetos se clasificó correctamente en las diferentes categorías de severidad de la AOS, expuestas anteriormente (Ecuación 4.10) (Grandini et al., 2020).

$$Acc\ por\ sujeto = \frac{VP+VN}{VP+VN+FP+FN} \tag{4.10}$$

Asimismo, se calculan el kappa y el ICC para comparar el AHI estimado con el AHI de referencia obtenido mediante la PSG. El **kappa de cohen** es una coeficiente que evalúa el nivel de concordancia entre dos clasificaciones, en este caso, el AHI estimado y el de referencia, omitiendo aquellos casos en los que coinciden por azar. Su expresión viene dada por la 'Ecuación 4.11' (Grandini et al., 2020).

$$k = \frac{p_0 - p_c}{1 - p_0} \tag{4.11}$$

En esta ecuación, p_0 representa la ratio de concordancia entre las clasificaciones reales y las obtenidas por la red neuronal, mientras que p_c representa la probabilidad de que coincidan por casualidad. El rango de valores posibles para este coeficiente oscila entre -1 y +1, donde un valor +1 representa una concordancia perfecta entre ambas clasificaciones, un kappa igual a -1 refleja que las predicciones nunca coinciden con la clasificación real, y un kappa igual a 0 indica que las coincidencias son únicamente debidas al azar (Grandini et al., 2020).

Finalmente, el ICC es un estadístico que sirve para valorar la fiabilidad de una medida. No obstante, hay diferentes modalidades de cálculo disponibles que pueden arrojar resultados diversos al ser aplicado sobre el mismo conjunto de datos. Por tanto, es importante definir la modalidad adecuada para cada aplicación. Cuando el objetivo es evaluar la fiabilidad de unos resultados con respecto a otros con las mismas características, como es el caso del AHI estimado y el AHI de referencia, la mejor opción es emplean el ICC (2, 1) según la convención de Shrout y Fleiss (Shrout & Fleiss, 1979). Su expresión viene dada por la 'Ecuación 4.12' (Koo & Li, 2016).

$$ICC = \frac{MSR - MSE}{(MSR + (k-1) * MSE + k * \frac{(MSC - MSE)}{n}} \tag{4.12}$$

MSR es el *mean square subject*, MSE el *mean square error* y MSC el *mean square colums.*

CAPÍTULO 5: RESULTADOS

En este capítulo se presentan los resultados obtenidos para la detección de eventos de apnea e hipopnea y la estimación del AHI empleando segmentos de 30 y 60 segundos de las señales de SpO_2, PR y su combinación SpO_2+PR, con o sin adyacencia. Además, en la última sección se incluyen los resultados de interpretación aplicando Grad-CAM.

5.1 RESULTADOS CON SEGMENTOS DE 30 SEGUNDOS

5.1.1 SIN ADYACENCIA

En primer lugar, se muestran los resultados obtenidos a partir de segmentos de 30 segundos sin considerar adyacencia. En la Tabla 5.1 se presentan los resultados tanto para la clasificación por segmento como para la clasificación por sujeto, considerando las diferentes señales de entrada: SpO_2, PR y SpO_2+PR con convoluciones 1D y 2D. El modelo con mejor rendimiento en ambas tareas es el que emplea exclusivamente la señal de SpO_2. En la Figura 5.1 se representan las matrices de confusión correspondientes a dicho modelo.

30-seg	ACC SEGMENTO	F1-SCORE	ACC SUJETO	KAPPA	ICC
SpO_2	**77.34%**	0.824	**64.51%**	0.507	0.829
PR	68.04%	0.755	32.90%	0.049	0.146
SpO_2+PR (1D)	75.53%	0.810	60.78%	0.456	0.793
SpO_2+PR (2D)	75.30%	0.809	59.48%	0.435	0.724

Tabla 5.1. Resultados con segmentos de 30 segundos sin adyacencia.

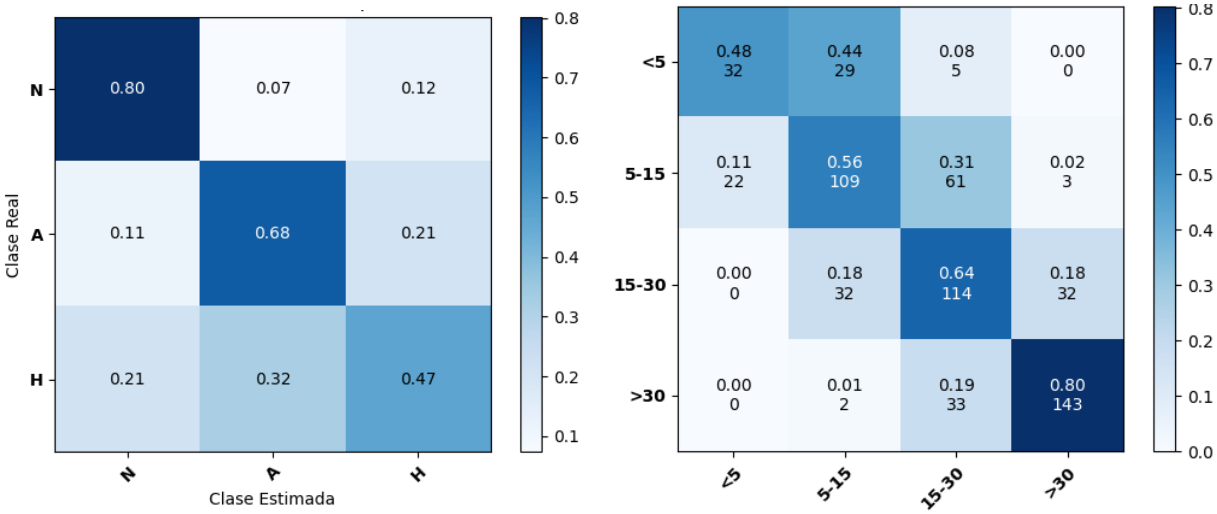

Figura 5.1 Matriz de confusión del modelo con mejor rendimiento en la clasificación por segmentos (izquierda) y sujeto (derecha) con la señal de SpO_2 en segmentos de 30 segundos sin adyacencia.

5.1.2 CON ADYACENCIA

A continuación, en la Tabla 5.2 se presentan los resultados obtenidos al utilizar un segmento de adyacencia, es decir, considerando el segmento previo y posterior al segmento correspondiente. Asimismo, en la Figura 5.2 se representa la matriz de confusión del modelo SpO_2+PR 2D con un segmento de adyacencia, que alcanza los mejores resultados en la clasificación por segmento, y la del modelo SpO_2, que obtiene el mejor desempeño en la clasificación por sujeto.

30-seg	ACC SEGMENTO	F1-SCORE	ACC SUJETO	KAPPA	ICC
SpO_2	77.58%	0.825	**65.96%**	0.528	0.829
PR	61.68%	0.707	34.28%	0.071	0.234
SpO_2+PR (1D)	75.91%	0.814	62.07%	0.476	0.811
SpO_2+PR (2D)	**81.28%**	0.850	65.32%	0.521	0.814

Tabla 5.2. Resultados con segmentos de 30 segundos y un segmento de adyacencia.

Por otra parte, también se han evaluado los modelos empleando dos segmentos de adyacencia. Esto implica la consideración de cinco segmentos de 30 segundos, es decir, una entrada a la red neuronal de 150 segundos. Para facilitar su evaluación, en la Tabla 5.3 se recogen los resultados obtenidos. La Figura 5.3 contiene las matrices de confusión de los modelos SpO_2+PR 2D y SpO_2 utilizando dos segmentos de adyacencia. Estos son los modelos con mejores resultados en la detección de eventos de apnea e hipopnea y estimación del AHI, respectivamente.

30-seg	ACC SEGMENTO	F1-SCORE	ACC SUJETO	KAPPA	ICC
SpO_2	80.67%	0.846	**69.85%**	0.582	0.858
PR	64.06%	0.725	34.68%	0.080	0.267
SpO_2+PR (1D)	80.99%	0.849	64.83%	0.513	0.836
SpO_2+PR (2D)	**85.13%**	0.876	67.59%	0.549	0.855

Tabla 5.3. Resultados con segmentos de 30 segundos y dos segmentos de adyacencia.

Por último, a modo de síntesis, en la Tabla 5.4 y en la Tabla 5.5 se recogen los modelos con mejor rendimiento en la clasificación por segmento y por sujeto, respectivamente, para facilitar su visualización y su posterior interpretación. Destaca la mejora en los resultados al incluir adyacencia.

	MODELO	ACC	F1-SCORE
Sin adyacencia	**SpO_2**	77.34%	0.824
Adyacencia de 1 segmento	**SpO_2+PR 2D**	81.28%	0.850
Adyacencia de 2 segmentos	**SpO_2+PR 2D**	85.13%	0.876

Tabla 5.4 Mejores resultados en la detección de eventos de apnea e hipopnea con segmentos de 30 segundos.

	MODELO	ACC	Kappa	ICC
Sin adyacencia	**SpO_2**	64.51%	0.507	0.829
Adyacencia de 1 segmento	**SpO_2**	65.96%	0.528	0.829
Adyacencia de 2 segmentos	**SpO_2**	69.85%	0.582	0.858

Tabla 5.5 Mejores resultados en la estimación del AHI con segmentos de 30 segundos.

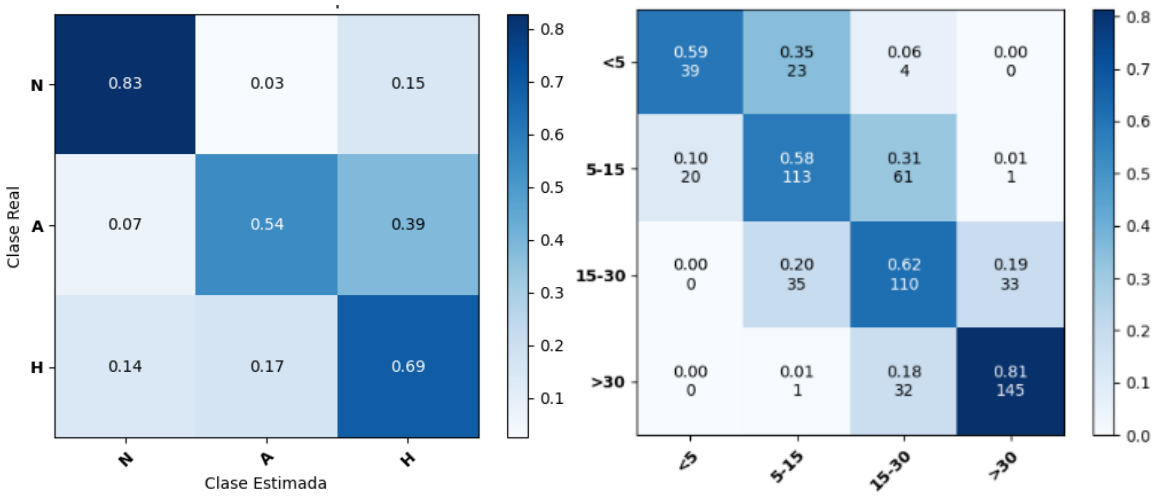

Figura 5.2 Matriz de confusión del modelo en la clasificación por segmentos (izquierda) con la señal de SpO_2+PR 2D con un segmento de adyacencia y matriz de confusión del modelo en la clasificación por sujeto (derecha) con la señal de SpO_2 en segmentos de 30 segundos con adyacencia de un segmento.

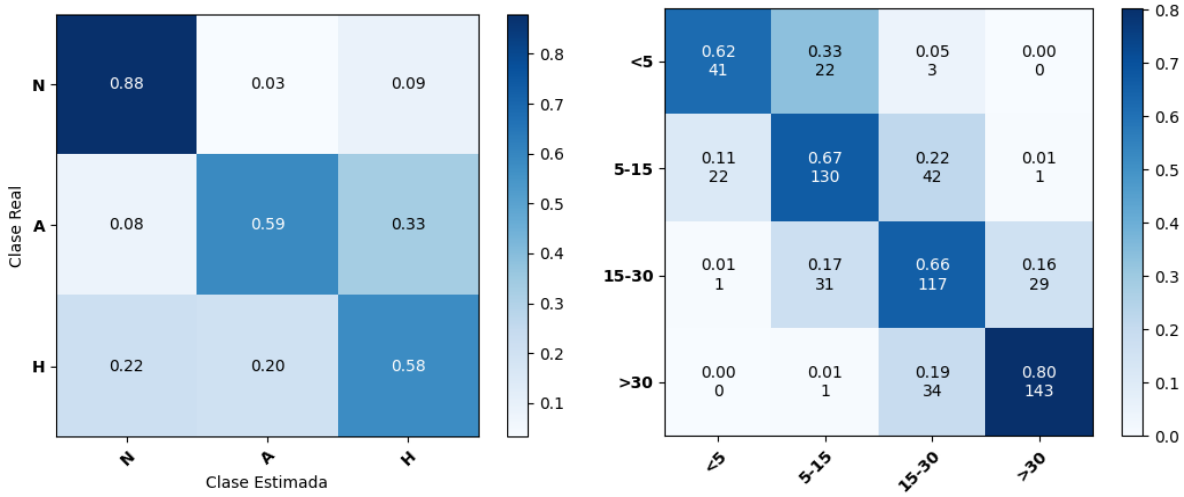

Figura 5.3 Matriz de confusión del modelo en la clasificación por segmentos (izquierda) con la señal de SpO_2+PR 2D con dos segmentos de adyacencia y matriz de confusión del modelo en la clasificación por sujeto (derecha) con la señal de SpO_2 en segmentos de 30 segundos con adyacencia de dos segmentos.

5.2 RESULTADOS CON SEGMENTOS DE 60 SEGUNDOS

5.2.1 SIN ADYACENCIA

Del mismo modo, la Tabla 5.6 recoge los resultados obtenidos al utilizar segmentos de 60 segundos sin adyacencia. Asimismo, en la Figura 5.4 se representan las matrices de confusión asociadas a los modelos que mostraron el mejor rendimiento en la clasificación por segmento y por sujeto. En el primer caso, se trata del modelo que fusiona las señales de SpO_2 y PR empleando convoluciones 1D. Por otro lado, en la estimación del AHI para clasificar los sujetos, el modelo que destaca utiliza la señal de SpO_2 con segmentos de 60 segundos sin adyacencia.

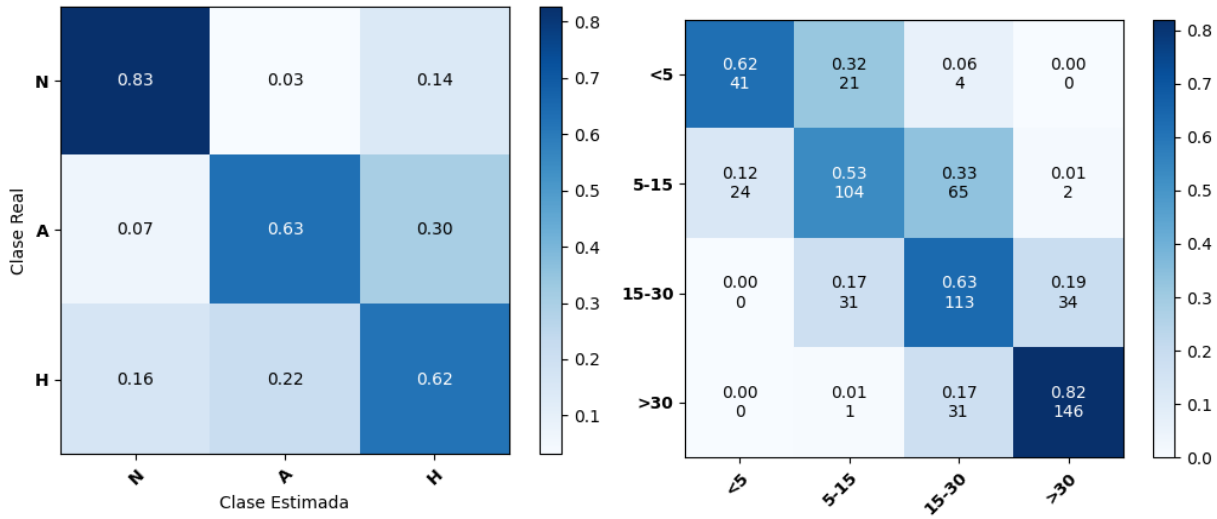

Figura 5.4 Matriz de confusión del modelo SpO_2+PR en la clasificación por segmentos utilizando segmentos de 60 segundos, sin adyacencia, y convoluciones 1D (izquierda). Matriz de confusión del modelo SpO_2 en la clasificación por sujeto empleando segmentos de 60 segundos sin adyacencia (derecha).

30-seg	ACC SEGMENTO	F1-SCORE	ACC SUJETO	KAPPA	ICC
SpO_2	78.13%	0.823	**65.48%**	0.523	0.813
PR	61.11%	0.678	34.04%	0.065	0.212
SpO_2+PR (1D)	**79.35%**	0.820	64.67%	0.513	0.824
SpO_2+PR (2D)	78.57%	0.815	63.21%	0.492	0.794

Tabla 5.6. Resultados con segmentos de 60 segundos sin adyacencia.

5.2.2 CON ADYACENCIA

A continuación, se presentan los resultados obtenidos mediante la estrategia de adyacencia. Para comenzar, se exploran los resultados al emplear un segmento de adyacencia, recogidos en la Tabla 5.7, lo cual equivale a un total de tres segmentos de 60 segundos como entrada: el segmento correspondiente, el previo y el posterior. Además, la Figura 5.5 incluye las matrices de confusión del modelo SpO_2 con un segmento de adyacencia. Este modelo alcanza los mejores resultados tanto en la detección de eventos de apnea e hipopnea para la clasificación por segmento como en la estimación del AHI para la clasificación por sujeto.

60-seg	ACC SEGMENTO	F1-SCORE	ACC SUJETO	KAPPA	ICC
SpO_2	**82.33%**	0.842	**68.40%**	0.564	0.846
PR	65.44%	0.711	35.66%	0.089	0.269
SpO_2+PR (1D)	80.80%	0.832	64.99%	0.516	0.806
SpO_2+PR (2D)	81.63%	0.837	64.34%	0.508	0.805

Tabla 5.7. Resultados con segmentos de 60 segundos y un segmento de adyacencia.

En segundo lugar, el empleo de dos segmentos de adyacencia implica utilizar cinco segmentos de 60 segundos con una longitud temporal de la entrada de 300 segundos. Con esta estrategia se consiguen los resultados que se muestran a continuación en la Tabla 5.8.

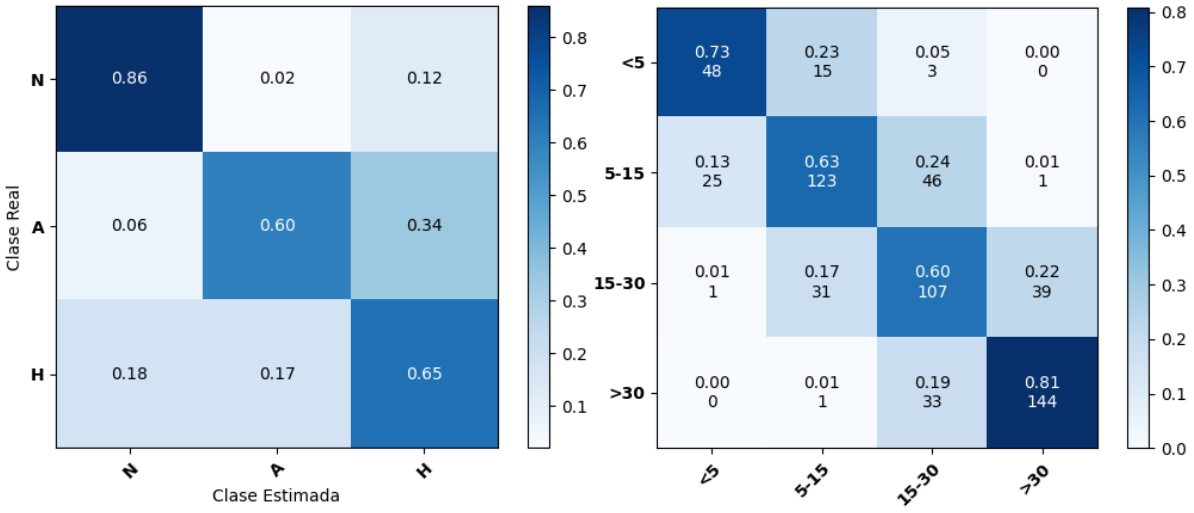

Figura 5.5 Matrices de confusión del modelo SpO_2 con segmentos de 60 segundos y un segmento de adyacencia en la clasificación por segmentos (izquierda) y sujeto (derecha).

60-seg	ACC SEGMENTO	F1-SCORE	ACC SUJETO	KAPPA	ICC
SpO_2	**83.53%**	0.850	**69.37%**	0.576	0.850
PR	67.42%	0.723	40.68%	0.171	0.391
SpO_2+PR (1D)	79.65%	0.824	66.29%	0.534	0.804
SpO_2+PR (2D)	78.89%	0.817	64.67%	0.511	0.751

Tabla 5.8. Resultados con segmentos de 60 segundos y dos segmentos de adyacencia.

Además, en la Figura 5.6 se representan gráficamente las matrices de confusión correspondientes al modelo que mostró el mejor rendimiento tanto en la clasificación por segmentos como en la clasificación por registro empleando dos segmentos de adyacencia. Concretamente, se trata del modelo con la señal de SpO_2 en segmentos de 60 segundos.

Finalmente, se han recogido los mejores modelos que emplean este tamaño de segmento en la detección de eventos de apnea e hipopnea y en la estimación del AHI en la Tabla 5.9 y la Tabla 5.10, respectivamente.

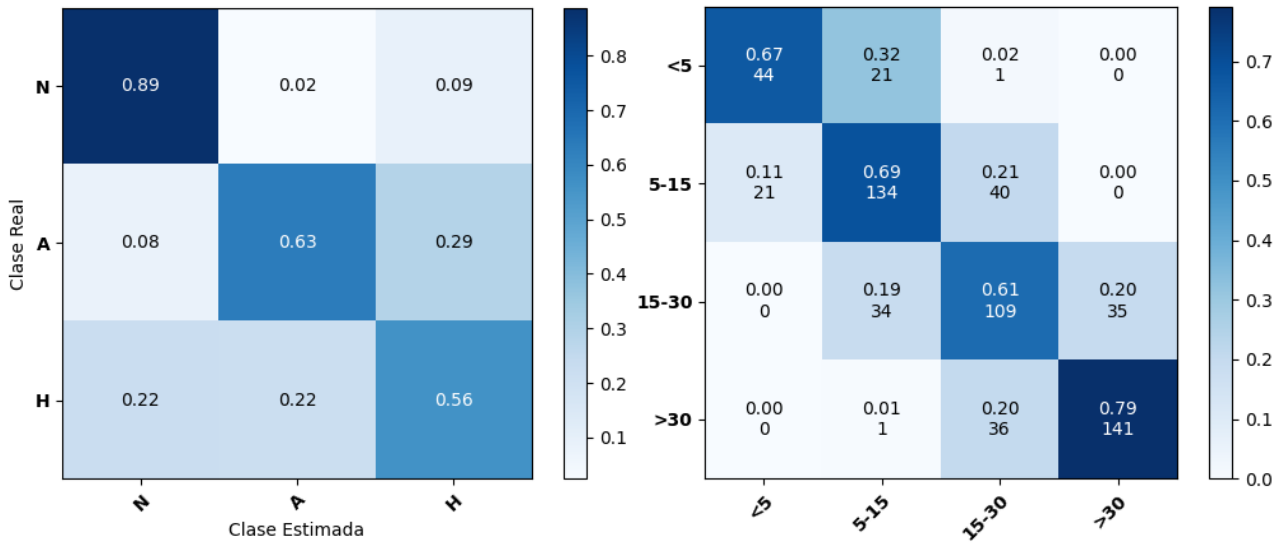

Figura 5.6 Matrices de confusión del modelo SpO_2 con segmentos de 60 segundos y un segmento de adyacencia en la clasificación por segmentos (izquierda) y sujeto (derecha).

	MODELO	ACC	F1-SCORE
Sin adyacencia	**SpO_2+PR 1D**	79.35%	0.820
Adyacencia de 1 segmento	**SpO_2**	82.33%	0.842
Adyacencia de 2 segmentos	**SpO_2**	83.53%	0.850

Tabla 5.9 Mejores resultados en la detección de eventos de apnea e hipopnea con segmentos de 60 segundos.

	MODELO	ACC	Kappa	ICC
Sin adyacencia	**SpO_2**	65.48%	0.523	0.813
Adyacencia de 1 segmento	**SpO_2**	68.40%	0.564	0.846
Adyacencia de 2 segmentos	**SpO_2**	69.37%	0.576	0.850

Tabla 5.10 Mejores resultados en la estimación del AHI con segmentos de 60 segundos.

5.3 RESULTADOS GRAD-CAM

En esta sección se presentan los resultados obtenidos a través de la aplicación de Grad-CAM sobre uno de los modelos que ha demostrado un buen rendimiento en la detección de eventos de apnea e hipopnea. En concreto, se trata del modelo que utiliza un tamaño de entrada de 60 segundos y emplea 2 segmentos de adyacencia. A partir de este modelo, se han generado los *heatmaps* correspondientes a los segmentos que han sido correctamente clasificados y que presentan la probabilidad más alta para cada una de las clases (N/A/H), descartando aquellos segmentos que contenían artefactos.

Para cada tipo de evento respiratorio, se han extraído los *heatmaps* utilizando la señal de SpO_2, la señal de PR y la combinación de señales de SpO_2 y PR. En cuanto a la combinación de ambas señales, se ha optado por emplear convoluciones 1D, ya que obtienen mejores resultados con un tamaño de entrada de 60 segundos y dos segmentos de adyacencia.

Por último, para garantizar la comparabilidad de los resultados para cada tipo de evento se realiza el siguiente procedimiento. En primer lugar, se seleccionan los segmentos que presentan la probabilidad más alta para cada evento específico con la señal de SpO_2. Después, a partir de estos segmentos se generan los *heatmaps* correspondientes utilizando las demás configuraciones de señales de entrada (PR y SpO_2+PR), aunque en algunos casos puede implicar la inclusión de segmentos clasificados incorrectamente. Sin embargo, esta estrategia garantiza la representación de los mismos segmentos en relación con cada uno de los eventos, lo cual facilita la interpretación de los resultados.

En las Figuras 5.7-5.9 se representan gráficamente los *heatmaps* de la clasificación de los eventos de apnea, hipopnea y respiración normal, respectivamente. A su vez, cada una de estas figuras se compone de tres *subplots*. El primero de ellos se corresponde con el mapa de calor de la señal de saturación de oxígeno en sangre, mientras que en el segundo *subplot* se representa el *heatmap* de la señal de frecuencia de pulso. Finalmente, el tercer *subplot* se corresponde con el mapa de calor de la combinación de ambas señales.

Adicionalmente, en cada figura se incluye la probabilidad de la predicción realizada por el modelo acompañada de la etiqueta correcta correspondiente al tipo de evento en cuestión. Asimismo, estos *heatmaps* están diseñados para resaltar las áreas con mayor importancia en la decisión del modelo, utilizando tonalidades de color rojo. Las regiones

con rojos más oscuros indican las zonas de la señal en las que el modelo se enfoca con mayor intensidad, evidenciando su relevancia en la clasificación. En cambio, a medida que la intensidad del rojo disminuye, se refleja una atención menos acentuada del modelo hacia esa región específica.

De forma general, se puede observar que las áreas de mayor relevancia para la detección de eventos de apnea o hipopnea se encuentran principalmente en aquellas zonas donde se producen descensos notables de la saturación de oxígeno en sangre y variaciones en la frecuencia cardiaca. Además, tienen especial relevancia los mínimos o *nadires* de la señal de SpO_2 y los descensos de la frecuencia cardiaca en la señal de PR. En cambio, para predecir las respiraciones normales, el modelo se centra en regiones estables de las señales de SpO_2 y PR, sin fluctuaciones bruscas o desaturaciones significativas.

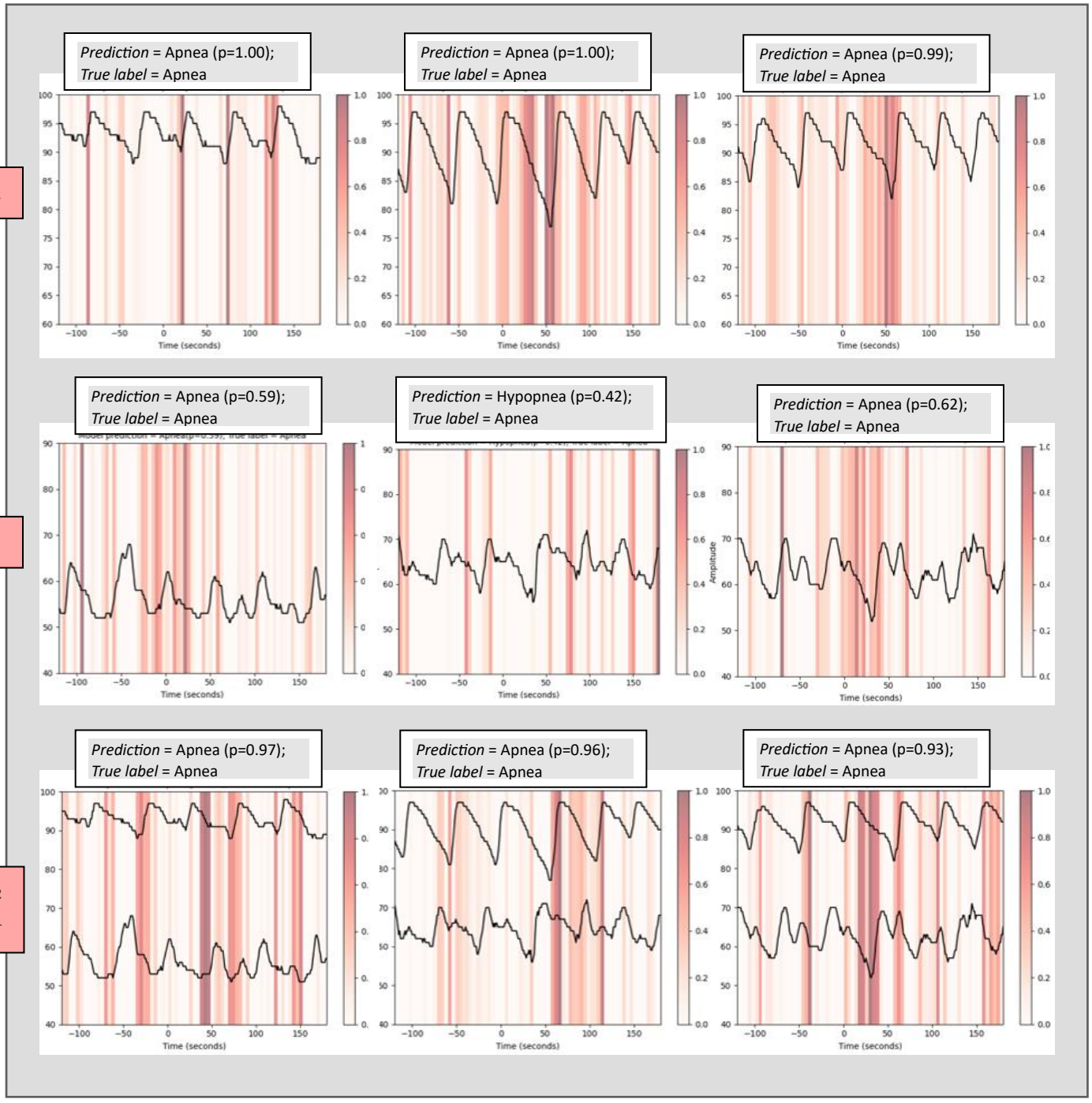

Figura 5.7. Heatmaps de la detección de eventos de apnea con las señales de SpO_2, PR y SpO_2+PR

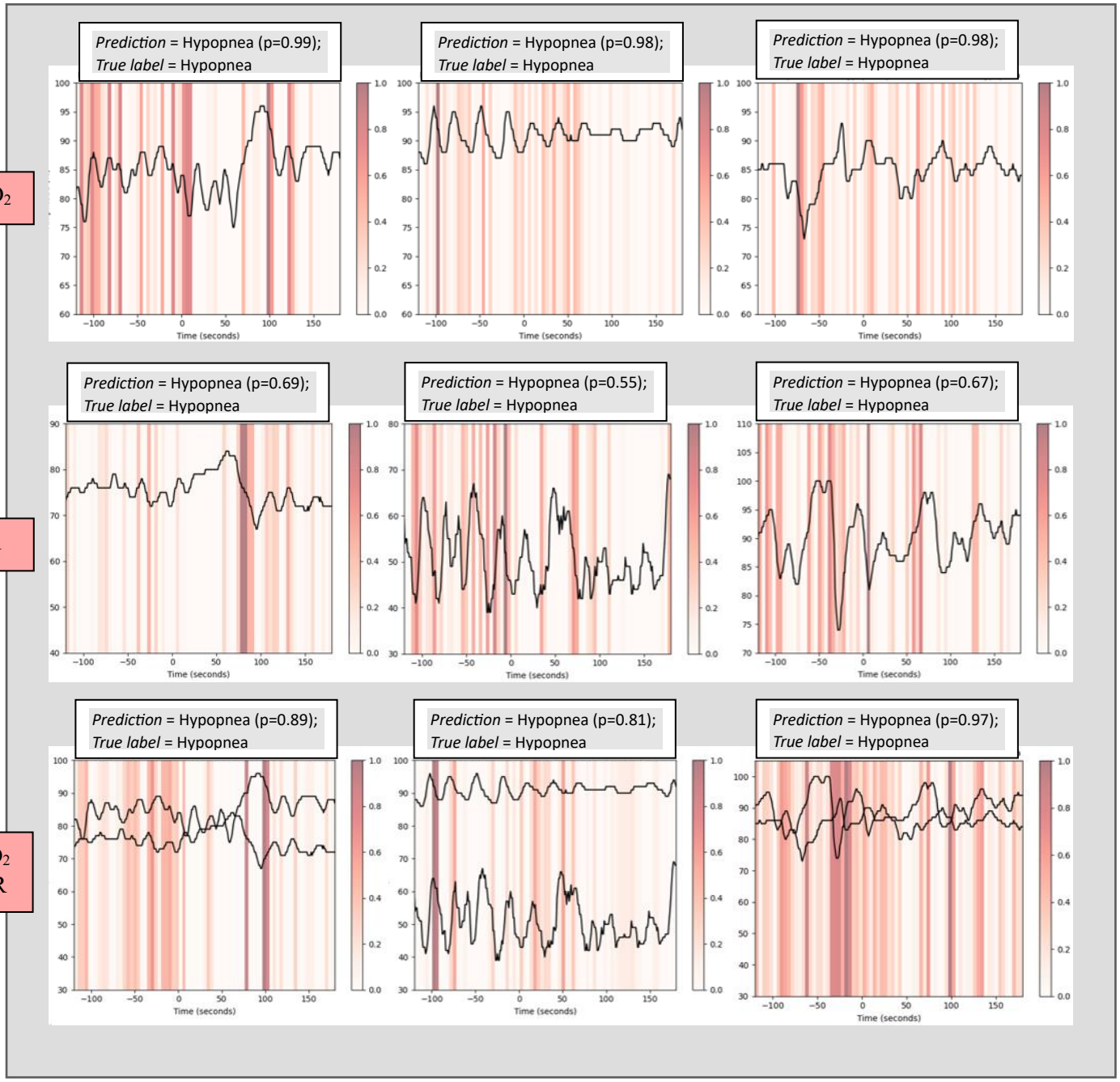

Figura 5.8. Heatmaps de la detección de eventos de hipopneas con las señales de SpO_2, PR y SpO_2+PR

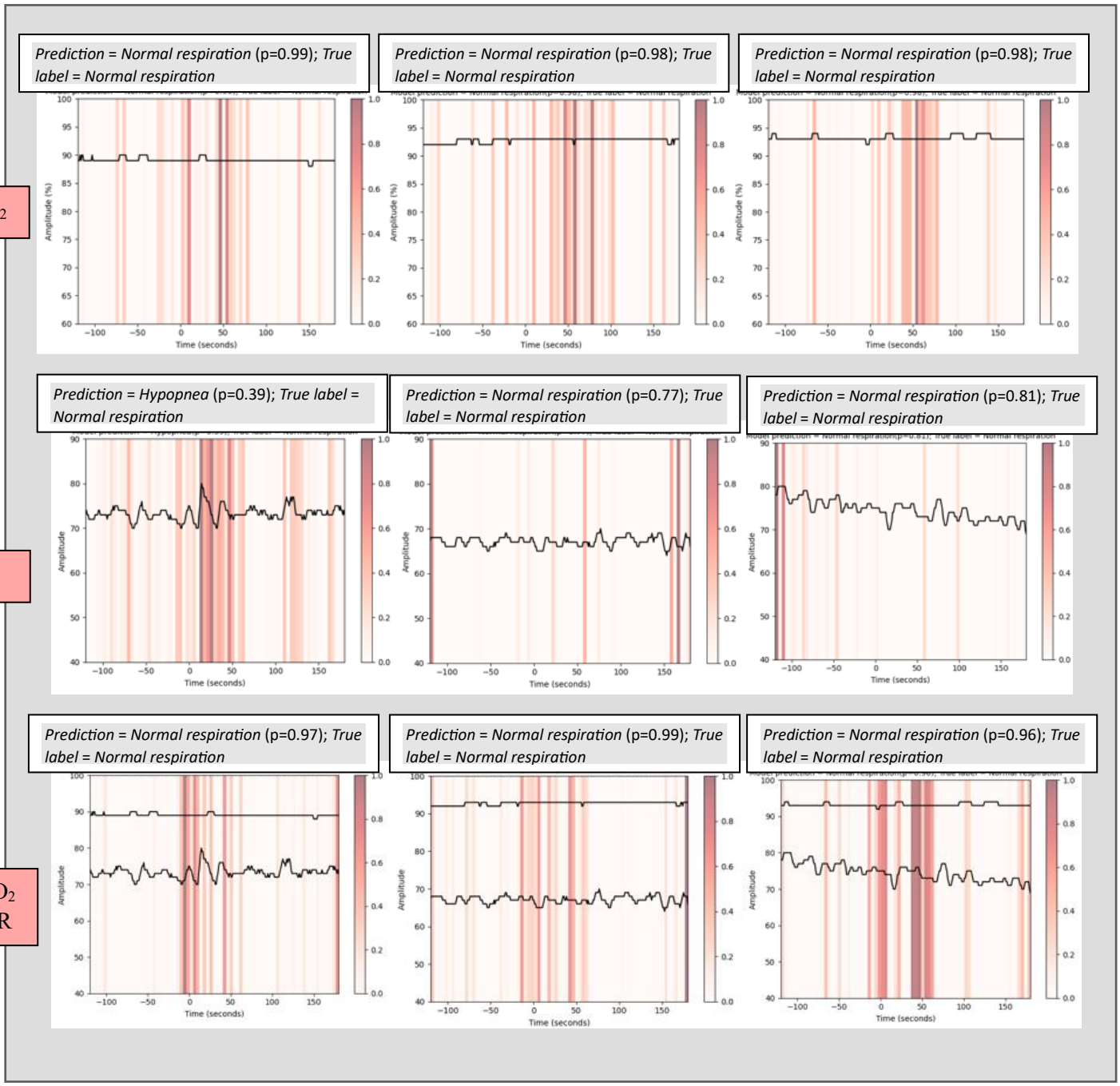

Figura 5.9. Heatmaps de la detección de eventos de respiraciones normales con las señales de SpO_2, PR y SpO_2+PR.

CAPÍTULO 6: DISCUSIÓN

En este sexto capítulo se analizan los resultados obtenidos en la clasificación de eventos de apnea e hipopnea, así como en la estimación del AHI. Asimismo, se lleva a cabo una comparación con los resultados de estudios previos, con el propósito de detallar los hallazgos obtenidos en este trabajo de investigación. Por último, se indican las principales limitaciones de este estudio.

6.1 CLASIFICACIÓN DE EVENTOS

6.1.1 RENDIMIENTO DE LAS ARQUITECTURAS CNN EN LA DETECCIÓN DE EVENTOS DE APNEA E HIPOPNEA

En primer lugar, comenzamos analizando los resultados sin la consideración de adyacencia para segmentos de entrada de 30 y 60 segundos. Es importante destacar que los resultados obtenidos en términos de exactitud con distintos tamaños de segmento no son directamente comparables, ya que el número de segmentos de cada clase depende del tamaño de segmento. Aun así, se puede observar en ambos casos como se obtiene mejores resultados con la señal de SpO_2 (exactitudes del 77.34% y 78.13% con 30 y 60 segundos, respectivamente) que con la señal de PR (exactitudes del 68.04% y 61.11% con 30 y 60 segundos, respectivamente). También es importante destacar que, siguiendo esta estrategia, el hecho de combinar ambas señales no conlleva mejoras notables en los resultados. Más aún, no se aprecian diferencias significativas al emplear convoluciones 1D o 2D. En este contexto, la utilización de la señal de SpO_2 con segmentos de 30 segundos resulta ser lo más efectivo en términos de exactitud (77.34%), mientras que con segmentos de 60 segundos los mejores resultados se consiguen combinando las señales de SpO_2 y PR y utilizando convoluciones 1D (79.35% de exactitud).

Por otro lado, se puede observar una tendencia muy interesante en los resultados para ambos tamaños de segmento: tanto la exactitud como el *F1-score* van mejorando a medida que aumenta la adyacencia, es decir, al ampliar la ventana temporal de entrada. Esto podría explicarse por el hecho de que un contexto temporal más extenso en torno al evento respiratorio contribuye a una mejor caracterización de dicho evento y, en consecuencia, a una detección más precisa. Respecto a las entradas de 30 segundos, se consiguen los mejores resultados al combinar ambas señales (SpO_2+PR) y al emplear convoluciones 2D, alcanzando una exactitud del 81.28% con un segmento de adyacencia y una exactitud del 85.13% con dos segmentos de adyacencia. En ambos casos, esto implica un aumento de la exactitud del 5% en comparación con el enfoque de convoluciones 1D. Respecto a las entradas de 60 segundos, al incorporar uno o dos segmentos de adyacencia, el modelo que emplea la señal de SpO_2 obtiene los mejores resultados. En particular, se alcanzan niveles de exactitud del 82.33% y 83.53%, respectivamente.

Considerando la estrategia de adyacencia, es importante destacar que la combinación de ambas señales no brinda mejoras notables en ninguno de los tamaños de segmento respecto a la utilización exclusiva de la señal de SpO_2. En el caso de segmentos de 30 segundos, los resultados mejoran ligeramente. Sin embargo, con segmentos de 60 segundos el rendimiento es peor al logrado con la señal de SpO_2. Esto sugiere la información proporcionada por la señal de PR no es complementaria a la de la señal de SpO_2 en la detección de eventos respiratorios.

En cuando a la discriminación de eventos respiratorios, la detección de la respiración normal es la que presenta el mejor rendimiento para ambos tamaños de segmento, con valores de exactitud que oscilan entre el 80% y 89%. En este grupo, la mayoría de los errores de clasificación tienden a confundirlo con hipopneas (9%-18%). Esto podría atribuirse a que los efectos fisiológicos de las hipopneas en las señales de SpO_2 y PR son menos evidentes en comparación con las apneas, lo que puede generar confusión con la respiración normal (Kulkas et al., 2017; Penzel, 2020, Chapter 15).

En lo que respecta a las apneas, se detectan correctamente en un rango del 54%-68%. Estos resultados sugieren que las apneas y las hipopneas presentan características distintas. Esto concuerda con los hallazgos de Kulkas et al. (2017), quienes informaron que las desaturaciones de oxígeno relacionadas a apneas obstructivas tienen una duración y profundidad significativamente mayores que las desaturaciones de SpO_2 relacionadas con hipopneas. En este contexto, las apneas mal detectadas podrían corresponder a eventos de corta duración que comparten similitudes con las hipopneas de larga duración.

Por último, las detección de hipopneas resulta ser la más desafiantes (47%-69%). Como se ha mencionado anteriormente, las hipopneas de larga duración pueden confundirse con apneas de corta duración (17-32%). Por otro lado, también se observa una confusión de las hipopneas con la respiración normal (14%-22%). Esta situación podría estar relacionada con hipopneas que están asociadas a eventos de *arousal* y que, en consecuencia, no producen un efecto fisiológico evidente en las señales de pulsioximetría (Kulkas et al., 2017; Penzel, 2020, Chapter 15).

En definitiva, los resultados obtenidos resaltan la importancia de la señal de entrada en el proceso de detección de apneas e hipopneas, reforzando el papel de la señal de SpO_2 en particular. Además, aumentar la ventana temporal en torno al evento respiratorio proporciona información adicional que contribuye a una mejor caracterización de dichos eventos. Por lo tanto, es interesante considerar tamaños de segmentos mayores o bien utilizar la estrategia de adyacencia.

6.1.2 INTERPRETACIÓN DE LAS DECISIONES TOMADAS POR LA CNN

Con respecto a la interpretación de las CNN vamos a analizar, a partir de los *heatmaps* obtenidos con Grad-CAM, el comportamiento del modelo que emplea segmentos de entrada de 60 segundos y dos segmentos de adyacencia, y así comprender los criterios adoptados para la toma de decisiones. Es importante destacar que en los *heatmaps* de Grad-CAM no se puede apreciar la contribución de cada una de las señales por separado. Po ellos, se optó por usar convoluciones 1D, ya que funcionan ligeramente mejor que las convoluciones 2D, de acuerdo con los resultados obtenidos.

En primer lugar, examinando los *heatmaps* de segmentos con **apnea**, se observa en la señal de SpO_2 la aparición de múltiples episodios de desaturación. En este caso, el comportamiento del modelo revela que concede especial importancia al punto mínimo de dichas desaturaciones, reflejándose mediante un color rojo intenso (ver Figura 5.7). Esto puede deberse a la mayor profundidad de las desaturaciones asociadas a una apnea (Kulkas et al., 2017; Penzel, 2020, Chapter 15). Por otro lado, empleando la señal de PR, los *heatmaps* dejan ver cierta variabilidad en la frecuencia cardiaca. En este caso, el modelo tiende a centrar su atención en puntos situados en la pendiente entre los máximos y mínimos relativos de la frecuencia cardiaca (ver Figura 5.7). Sin embargo, siguiendo este criterio, se observa que las probabilidades de clasificación no son especialmente altas y en algunos casos, las clasificaciones son erróneas.

En cuanto a la combinación de ambas señales, las regiones de interés varían en comparación con las observadas al considerar las señales de forma individual. En el primer gráfico, el modelo se centra en una región con una desaturación acompañada de un aumento en la frecuencia cardiaca (taquicardia), situación similar a la que se refleja en el tercer *subplot*. En cambio, en el segundo *subplot*, la atención del modelo se enfoca en una zona donde se produce una resaturación acompañada con una disminución de la frecuencia cardiaca (bradicardia). Asimismo, al fusionar las señales se aprecia una disminución de las probabilidades de salida de la CNN en comparación con el uso exclusivo de la señal de SpO_2. Esto sugiere que considerar el valor mínimo de la desaturación contribuye favorablemente a la discriminación de las apneas. Sin embargo, al incorporar la señal de PR el modelo tiende a desviar su atención de estos mínimos hacia regiones de la pendiente, influido por el comportamiento de dicha señal.

Con respecto a la detección de **hipopneas**, en los *heatmaps* correspondientes a la señal de SpO_2 se observa que el modelo muestra una tendencia a enfocarse en el periodo comprendido entre el comienzo de la desaturación y el *nadir* (donde se da el valor mínimo). Esto puede estar relacionado con la menor duración y profundidad de las desaturaciones asociadas a hipopnea (Kulkas et al., 2017; Penzel, 2020, Chapter 15) . Asimismo, se puede apreciar que, para la discriminación de hipopneas, el modelo emplea principalmente la información de dos segmentos anteriores al segmento en cuestión. Esto puede deberse a que compara el mínimo y profundidad de la desaturación con las de desaturaciones que ocurren en segmentos adyacentes. Esto difiere del comportamiento observado con las apneas, donde se consideraba principalmente la información del segmento central. Esta observación va en concordancia con lo comentado anteriormente, ya que los resultados reflejaban que la discriminación de las hipopneas mejoraba considerablemente conforme aumentaba la adyacencia.

Por otro lado, al analizar los *heatmaps* generados con la señal de PR se ve que las regiones de interés del modelo no están claramente definidas (ver Figura 5.8). En algunos casos se toma información tanto de segmentos previos como posteriores al segmento en cuestión, mientras que en otras situaciones el modelo centra toda su atención en una región en particular, por ejemplo, un descenso pronunciado de la frecuencia cardiaca. En definitiva, no se puede apreciar un criterio único para clasificar un segmento como 'hipopnea' utilizando esta señal. Esto concuerda con el bajo rendimiento de la señal de PR a la hora de detectar apneas e hipopneas.

A su vez, al combinar ambas señales se observa que el modelo desplaza su atención hacia las regiones en las que se produce el mayor descenso de la señal de SpO_2 acompañado de un aumento de la frecuencia cardiaca. Este comportamiento es similar al observado en la discriminación de las apneas. Como ya se ha obtenido con la señal de SpO_2, la principal diferencia en la forma de actuar del modelo radica en el mayor énfasis del modelo en la información previa al evento, por lo que cobra mayor importancia la estrategia de adyacencia. Sin embargo, esta información previa puede variar considerablemente entre diferentes segmentos, lo que dificulta la predicción de esta clase. En consecuencia, los casos en los que el modelo confunde la clase 'hipopnea' con la clase 'apnea' pueden deberse a que se produce un descenso profundo en la saturación de oxígeno en sangre.

Por último, en los *heatmaps* correspondientes a los segmentos de **respiraciones normales** se observa que, en el caso de la señal de SpO_2, el modelo dirige su atención a regiones donde la señal se mantiene constante. Esto es coherente con el comportamiento de dicha señal, ya que en condiciones normales es muy estable. Del mismo modo, cuando no se producen eventos respiratorios, la señal de PR presenta una variabilidad pequeña, sin apenas aumentos o descensos notables de la frecuencia cardiaca. No obstante, puede haber situaciones en las cuales varíe la frecuencia por razones no relacionadas con eventos de apnea o hipopnea. Concretamente, en uno de los ejemplos representados se aprecia un incremento notable de la frecuencia cardiaca, el cual el modelo ha clasificado erróneamente como hipopnea. Sin embargo, durante este tiempo, la señal de SpO_2 se ha mantenido constante. Esta situación nuevamente resalta la falta de fiabilidad de los resultados obtenidos con la señal de PR. Finalmente, al combinar ambas señales se puede apreciar que los tramos constantes de las señales cobran especial relevancia en la toma de decisiones. Asimismo, se ve que el modelo tiende a enfocarse en la información del segmento central, lo que indica que, en el caso particular de los segmentos con respiraciones normales, la adyacencia no contribuye a la mejora de los resultados.

6.2 ESTIMACIÓN DEL AHI

La estimación de AHI es necesaria para clasificar los sujetos en cuatro grupos distintos en función de la severidad de la AOS. En este contexto, se ha requerido la implementación de una regresión lineal para tratar de obtener una estimación más robusta y confiable del AHI, minimizando la influencia de los valores atípicos.

Al examinar los resultados presentados anteriormente, se puede observar que, tanto para segmentos de entrada de 30 segundos como de 60 segundos, los mejores resultados se logran al emplear exclusivamente la señal de SpO_2 independientemente de la estrategia de adyacencia utilizada. Además, se aprecia una tendencia similar a la observada en la detección de eventos de apnea e hipopnea: los resultados mejoran conforme aumenta la adyacencia considerada y el rendimiento de los modelos con ambos tamaños de segmento se vuelve prácticamente similar. Concretamente, al utilizar dos segmentos de adyacencia con la señal de SpO_2 se alcanzan los mejores resultados, logrando una exactitud del 69.85% y un kappa de 0.582 para segmentos de 30 segundos, así como una exactitud del 69.37% y un kappa de 0.576 para segmentos de 60 segundos. Sin embargo, es importante

destacar que, cuando no se utiliza adyacencia, el modelo que emplea segmentos de 60 segundos supera al de 30 en términos de exactitud.

En cambio, la señal de PR no muestra un rendimiento favorable en la estimación del AHI. La exactitud más alta obtenida es del 40.68%, lograda por el modelo que emplea entradas de 60 segundos y dos segmentos de adyacencia. Estos resultados van en la línea del rendimiento observado previamente con esta misma señal en la detección de eventos de apnea e hipopnea. No obstante, a pesar de su bajo rendimiento, se puede apreciar una mejora en la estimación del AHI conforme se aumenta la adyacencia.

Además, al combinar ambas señales, los resultados experimentan una ligera disminución en comparación con la utilización exclusiva de la señal de SpO_2. En este caso, la inclusión de la señal de PR perjudica más de los que beneficia en términos de exactitud en la estimación del AHI. Esto sugiere que la señal de PR no proporciona información complementaria a la de SpO_2 a la hora de diagnosticar la AOS.

Profundizando en los modelos que hacen uso de la señal de SpO_2, se observa que la mayoría de los sujetos erróneamente clasificados son asignados a un grupo de severidad contiguo. Por ejemplo, los individuos con un AHI superior a 30 se clasifican correctamente como AOS grave (en torno al 80%) o erróneamente AOS moderado (en torno al 20%), mientras que los sujetos con no AOS tienden a ser clasificados correctamente (62%-73%) o como AOS leve (23%-32%), con muy pocos sujetos siendo erróneamente clasificados como AOS moderado (2%-6%). Asimismo, en los pacientes con AOS moderado, los errores ocurren principalmente con los grupos AOS leve y AOS grave, mientras que los adultos con AOS leve erróneamente diagnosticados por la CNN son asignados a los grupos AOS moderado y no AOS.

Por otra parte, al analizar la utilización de la señal de PR para tratar de explicar los resultados desfavorables en la estimación del AHI, se observa que el origen se debe a una tendencia del modelo a clasificar a la gran mayoría de los pacientes (más del 70%) en la categoría ‘AOS moderado’. Además, entre los sujetos restantes, una proporción significativa es incluida en la categoría ‘AOS leve’. Esto resulta intrigante puesto que el modelo prácticamente no clasifica a ningún sujeto en el grupo ‘No AOS’ ni en el de ‘AOS grave’. De hecho, en el caso de segmentos de 30 segundos sin adyacencia, no se observa ningún individuo que sea clasificado como ‘No AOS’.

Finalmente, al combinar ambas señales se observan unas tendencias similares a las de la señal de SpO_2 en solitario, pero con unos valores de exactitud y de kappa inferiores. Asimismo, también se observa la consideración de adyacencia contribuye a mejorar la exactitud de la clasificación del modelo, independientemente de la señal utilizada. Esto es porque proporciona un contexto temporal que permite mejorar la detección de los eventos respiratorios, lo que a su vez se traduce en estimaciones del AHI más confiables.

6.3 COMPARACIÓN CON OTROS ESTUDIOS

La comparación entre estudios que llevan a cabo la tarea de detección automática de eventos de apnea e hipopnea no es sencilla, ya que las diversas investigaciones difieren

en términos de la base de datos empleada, el tipo de señal utilizada, el tamaño de los segmentos y, además, en el enfoque de *deep learning* utilizado.

De todos los artículos revisados, únicamente dos de ellos emplean la señal de SpO_2 junto con una CNN para abordar esta tarea. Concretamente, se trata de las investigaciones llevadas a cabo por Huttunen et al. (2023). y Mostafa et al. (2020). En la Tabla 6.1 se recogen los resultados obtenidos en estos estudios junto con los conseguidos en este TFG, con el fin de facilitar la comparación y la evaluación de las contribuciones de cada uno de ellos.

AUTOR	**SEÑAL**	**BASE DE DATOS**	**METODOLOGÍA**				**RESULTADOS**
			MÉTODO	TAMAÑO	ETIQUETA	AHI	
(Huttunen et al., 2023)	PPG y SpO_2*	Privada (877 sujetos)	U-time	1s	N/A/H	Si	**Sujeto**: ICC (0.946), kappa (0.54)
(Mostafa et al., 2020)	SpO_2	HuGCDN2008 (40 sujetos) //Apnea-ECG (8 sujetos) //UCD (25 sujetos)	CNN	1 min// 3 min// 5 min	N/A	No	**Segmento**: AED 1 min: Acc (94.24%), Se (92.04%), Sp(95.78%); AED 3 min: Acc (93.93%), Se (89.87%), Sp (96.78%); UCD 1 min: Acc (84.85), Se (58.32%), Sp (93.32%); UCD 3 min: Acc (85,79%), Se (60.38%), Sp (93.9%)
Este TFG	SpO_2	MESA (2056 sujetos)	CNN	1 min// 3 min	N/A/H	Si	**Segmento:** 30 segundos**:** Acc **(**85.13%); 60 segundos: Acc (83.53%) **Sujeto**: 30 segundos: kappa (0.582); 60 segundos: kappa (0.576)

*Tabla 6.1. Comparación de las investigaciones que emplean la señal de SpO_2. * La señal de SpO_2 y PPG se utiliza en el modelo 1. Siglas: normal (N), apnea (A), hipopnea (H), accuracy (acc), intraclass correlation coeficient (ICC)), convolutional neural network (CNN), apnea-ECG database (AED).*

En primer lugar, Huttunen et al. (2023) diferencian entre respiraciones normales, apneas e hipopneas, aunque su enfoque no incluye una detección de eventos por segmento. En lugar de ello, realizan una clasificación por sujeto mediante la estimación del AHI, obteniendo un kappa de 0.540. De forma comparativa, en este TFG, el modelo que utiliza la señal de SpO_2 con segmentos de 30 segundos alcanza un kappa de 0.582, mientras que al emplear segmentos de 60 segundos se logra un kappa de 0.576. En ambos casos, se consideró una adyacencia de dos segmentos. Por lo tanto, nuestros resultados en la clasificación por sujeto superan ligeramente los obtenidos en investigaciones previas que emplearon la misma señal.

Por otro lado, Mostafa et al. (2020) sí que llevan a cabo detección de eventos utilizando dos bases de datos distintas. Por ejemplo, con Apnea-ECG y segmentos de un minuto logran una exactitud del 94.04%. Sin embargo, es importante destacar que esta base de datos solo cuenta con 8 registros que incluyen la señal de SpO_2, lo cual limita la fiabilidad de los resultados al tratarse de una muestra muy reducida. En cambio, al emplear la base de UCD, que contiene un mayor número de registros, logran una exactitud del 84.85% con segmentos de 1 minuto. No obstante, cabe mencionar que su investigación se limita a discriminar entre respiraciones normales y apneas, mientras que en este TFG se realiza una clasificación multiclase que distingue entre apneas e hipopneas, logrando resultados prácticamente equivalentes en términos de exactitud. Concretamente, los modelos que

emplean la señal de SpO_2 con dos segmentos de adyacencia y tamaños de entrada de 30 y 60 segundos logran una exactitud del 85.13% y 83.53%, respectivamente.

Asimismo, en varios de los artículos revisados combinan múltiples señales respiratorias, obteniendo resultados que indican que la fusión de señales favorece la detección de eventos de apnea e hipopnea. En este TFG, al considerar las señales de pulsioximetria se observan algunos casos concretos, en la clasificación por segmento, en los que se producen mejoras respecto de la utilización en solitario de la señal de SpO_2. Estos casos incluyen los modelos con entradas de 30 segundos y adyacencia, y el modelo con entradas de 60 segundos sin adyacencia. No obstante, para los modelos restantes y en lo que respecta a la clasificación por sujeto, la inclusión de la señal de PR no resulta en una mejora del rendimiento. Esto se debe a que la información proporcionada por esta señal de la base de datos MESA no es complementaría a la proporcionada por la de SpO_2 a la hora de caracterizar los eventos de apnea e hipopnea. Sin embargo, estos resultados no pueden ser comparados con investigaciones anteriores, ya que ninguno de los artículos incluidos en la revisión bibliográfica utilizó la señal de PR para la detección de eventos de apnea e hipopnea en adultos. En cambio, utilizan la señal de SpO_2 junto con otras señales respiratorias como el flujo aéreo o los movimientos torácicos y abdominales.

Finalmente, tras comparar los resultados con los de otras investigaciones previas, se puede afirmar que este TFG es el primero en emplear las señales de pulsioximetría para la detección de eventos de apnea e hipopnea, llevando a cabo una clasificación de tres clases (N/A/H). Es importante destacar que Huttunen et al. (2023) también realizan una distinción entre apneas e hipopneas en su investigación. La principal diferencia radica en que, mientras ellos utilizan las señales de SpO_2 y PPG, en este TFG se ha optado por emplear las señales de SpO_2 y PR. La ventaja de utilizar estas señales derivadas de la pulsioximetría es que son más accesibles y pueden ser almacenadas por los pulsioxímetros portátiles. Sin embargo, la señal de PPG no se almacena de manera estándar en estos dispositivos, lo cual limita su aplicabilidad (*@ Www.Nonin.Com*; *@ Www.Masimo.Com*). Asimismo, este TFG logra superar los resultados presentes en la literatura en cuanto a la estimación del AHI empleando la señal de SpO_2. Por último, resaltar que este trabajo de investigación es el único entre todos los artículos revisados que incorpora técnicas de XAI, con el fin de explicar las decisiones del modelo y favorecer la interpretación de los resultados obtenidos.

6.4 LIMITACIONES

Una de las principales limitaciones enfrentadas durante la realización de este TFG fue el desequilibrio de clases presente en los registros de la base de datos de MESA. La gran mayoría de eventos respiratorios correspondía a la categoría de ‘respiraciones normales’, mientras que las clases ‘apnea’ e ‘hipopnea’ contaban con un número mucho menor de muestras. Para abordar este problema, fue necesario aplicar técnicas de *oversampling*, que consistieron en aumentar la cantidad de muestras en las clases minoritarias. Sin embargo, el incremento en el número de datos resultó en un aumento considerable del coste computacional para entrenar y evaluar los diversos modelos, así como para probar diferentes arquitecturas. En consecuencia, no fue posible disponer del tiempo necesario para implementar otras metodologías de *deep learning*, como las LSTM o los

Transformer, para comparar con el rendimiento de la CNN en la tarea de detección de eventos de apnea e hipopnea.

Otra de las limitaciones encontradas está relacionada con la detección de eventos de apnea e hipopnea, ya que se han podido clasificar correctamente los eventos respiratorios, pero no se ha identificado su inicio ni su duración. Esto se debe a la presencia de un retardo variable entre el comienzo de la apnea o hipopnea y la aparición de desaturaciones en la señal de SpO_2 o patrones de bradicardia/taquicardia en la señal de PR asociados al evento (Kulkas et al., 2013). Por lo tanto, esta variabilidad temporal dificulta la caracterización exacta de dichos eventos.

También es importante destacar que el enfoque de interpretabilidad y visualización implementado, basado en Grad-CAM, no es la única forma de realizar XAI. A pesar de que se ha demostrado la utilidad de los *heatmaps* generados por Grad-CAM para identificar patrones en las señales de SpO_2 y PR, también existen otras técnicas de XAI que podrían proporcionar interpretaciones alternativas, cuya elección dependerá del contexto y los objetivos de la investigación. Además, Grad-CAM ha sido diseñado específicamente para su aplicación en CNNs, lo cual limita su aplicabilidad a otros tipos de modelos de DL.

Por otro lado, esta metodología no ha sido evaluada en subgrupos específicos de población, como rangos de edad, género o grupos de índice de masa corporal, entre otros. Esto implica que, aunque se han alcanzado unos buenos resultados en la detección de eventos de apnea e hipopnea, no se puede determinar en qué subgrupos de AOS es más apropiada esta metodología aplicada a las señales de pulsioximetría.

Finalmente, otra de las limitaciones está relacionadas con el tamaño de la base de datos utilizada. Aunque MESA es relativamente grande en comparación con las bases de datos empleadas en otros estudios, con un total de 2056 sujetos que han sido registrados con pulsioximetría, sería necesario disponer de conjuntos de datos adicionales con un mayor número de sujetos para poder evaluar la capacidad de generalización de los modelos implementados.

CAPÍTULO 7: CONCLUSIONES Y LÍNEAS FUTURAS

A lo largo de este TFG, se ha explorado el uso de las CNN como una metodología para la detección automática de eventos de apnea e hipopnea, evaluando su rendimiento en los registros de sujetos adultos de la base de datos MESA. Para finalizar este trabajo de investigación, una vez alcanzados los principales objetivos, en este último capítulo se detallan las contribuciones realizadas, las conclusiones extraídas y varias líneas de investigación que podrían seguirse en trabajos futuros.

7.1 CONTRIBUCIONES

En este trabajo de investigación se llevan a cabo una serie contribuciones en el contexto de la AOS para la detección automática de eventos de apnea e hipopnea. Estas aportaciones pueden resumirse en tres puntos principales, los cuales se describen a continuación.

- En primer lugar, este trabajo es pionero en la detección automática de eventos de apnea e hipopnea basada en las señales de SpO_2 y PR en adultos. Para ello, se lleva a cabo una clasificación multiclase distinguiendo entre respiración normal, apneas e hipopneas. Además, se estima el AHI para clasificar los sujetos en función de la severidad de la AOS, lo que enriquece la metodología desarrollada y amplía su aplicabilidad clínica.

- En segundo lugar, se lleva a cabo una comparación del rendimiento obtenido mediante diversas configuraciones de señales. En concreto, se evalúan las señales de SpO_2, PR y una combinación de ambas (SpO_2+PR) en el contexto de la detección de apneas e hipopneas, así como en la estimación del AHI. Además, se exploran diferentes tamaños de segmentos (30 y 60 segundos), así como la estrategia de adyacencia (0,1 y 2 segmentos a cada lado).

- Además, este estudio destaca por ser el único en la literatura actual en el contexto de la detección automática de eventos de apnea e hipopnea que incorpora técnicas de XAI. Estas técnicas se emplean con el objetivo de identificar patrones en el comportamiento de las señales, tratar de explicar las decisiones tomadas por el modelo y facilitar la interpretación de los resultados. Por lo tanto, esta contribución tiene un gran potencial para mejorar la confianza en los resultados obtenidos, lo cual es especialmente importante para incorporar nuevos métodos automáticos de diagnóstico en el ámbito de la salud y en el contexto de la AOS.

7.2 CONCLUSIONES

Las conclusiones extraídas a partir de los resultados de este trabajo de investigación se enumeran a continuación.

- Los métodos de *deep learning* son de gran utilidad en la detección automática de eventos de apnea e hipopnea a partir de señales de pulsioximetría, así como la estimación del AHI. En particular, este trabajo ha demostrado que las CNN constituyen un algoritmo muy adecuado para llevar a cabo dicha tarea, eliminando la necesidad de implementar etapas previas de extracción y selección de características. Esto presenta una alternativa con gran potencial para simplificar el diagnóstico de la AOS, al mismo tiempo que aporta objetividad y eficiencia en términos de tiempo y recursos económicos.

- Se ha demostrado la utilidad de las señales de pulsioximetría en la detección de eventos de apnea e hipopnea, con valores de exactitud del 85.13% y 83.53% y valores de F1-*score* de 0.876 y 0.850 para segmentos de 30 y 60 segundos, respectivamente. Además, en este estudio también se ha demostrados que la señal de SpO_2 contiene la información necesaria para realizar una estimación precisa del AHI. Los valores de kappa obtenidos para segmentos de 30 (0.582) y 60 segundos (0.570) confirman la capacidad de esta señal para proporcionar estimaciones fiables.

- La señal de PR de la base de datos MESA no ha demostrado ser útil por si sola para la detección de eventos de apnea e hipopnea. Además, al fusionarse con la señal de SpO_2 (SpO_2+PR) no se produce una mejora significativa en la detección de eventos respiratorios y en la estimación del AHI.

- A medida que se incrementa la adyacencia considerada, se ha observado una mejora progresiva en los resultados tanto para la detección de eventos como para la estimación del AHI. El hecho de ampliar la ventana temporal de entrada conlleva a una caracterización más completa de los eventos de apnea e hipopnea, permitiendo capturar relaciones y patrones que podrían no ser evidentes en segmentos más cortos.

- La clasificación de hipopneas presentan el mayor desafío entre los diferentes eventos respiratorios. En su detección, tienden a ser confundidas principalmente con apneas, pero también muestran cierta tendencia a clasificarse erróneamente como respiraciones normales. Estas dificultades en la discriminación se deben a las características a veces ambiguas de las hipopneas en comparación con las apneas y la respiración normal. Sin embargo, se ha observado que al aumentar la adyacencia en los segmentos de entrada se mejoran los resultados. Esto es debido a que, tal y como ve en los *heatmaps*, la información temporal previa al vento adquiere una gran importancia.

- La inclusión de técnicas XAI facilita la comprensión de los resultados, así como la comprensión del razonamiento detrás de las decisiones del modelo en la detección de eventos de apnea e hipopnea, mejorando la transparencia e

interpretabilidad de los modelos de *deep learning*. Esto es de vital importancia, especialmente de cara a ser utilizado en la práctica clínica.

En conclusión, este trabajo de investigación ha contribuido significativamente en el campo de la detección automática de eventos de apnea e hipopnea. Se ha demostrado la utilidad de las técnicas de *deep learning*, especialmente las CNN, en la clasificación precisa de estos eventos a partir de la señal de SpO_2 y PR. Además, la estrategia de adyacencia ha demostrado ser importante para mejorar la detección de las hipopneas, resaltando la importancia de la información temporal previa. Por último, la implementación de Grad-CAM para obtener los *heatmaps* ha proporcionado una justificación de las decisiones de los modelos de *deep learning*, facilitando la interpretación de los resultados.

7.3 LÍNEAS FUTURAS

Durante el desarrollo de este TFG, se han identificado varias áreas de interés que pueden ser objeto de un análisis más profundo en futuras investigaciones:

- En primer lugar, sería interesante explorar otras metodologías presentes en la literatura científica, como, por ejemplo, la combinación de arquitecturas como CNN + RNN o CNN + *Transformer*, Estos enfoques, que han demostrado un buen rendimiento en la detección de eventos con otras señales cardiorrespiratorias, podrían proporcionar nuevas perspectivas en la detección de apneas e hipopneas en adultos mediante la señal de pulsioximetría.

- Asimismo, sería interesante ampliar la aplicación de técnicas de XAI, como Grad-CAM, para mejorar la comprensión de los modelos. En lugar de aplicarlo únicamente en la última capa convolucional, podría ser interesante extenderlo al resto de capas, profundizando así en la visualización de los patrones de activación y su relación con las decisiones del modelo. En este sentido, también sería interesante evaluar otras técnicas de XAI como SHAP.

- La evaluación de nuestras metodologías de *deep learning* en subgrupos de población que presentan diferentes características clínicas ayudaría identificar aquellos fenotipos dentro de la AOS en los que los enfoques basados en señales de oximetría logran un rendimiento superior.

- Por último, sería conveniente probar el modelo desarrollado en otras bases de datos diferentes que incluyan registros con señales de pulsioximetría. Esto permitiría determinar si la limitada utilidad de la señal de PR observada en este trabajo es una particularidad de la base de datos empleada. Además, permitiría valorar la capacidad de generalización del modelo ante nuevos conjuntos de datos.

REFERENCIAS

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Almutairi, H., Hassan, G. M., & Datta, A. (2021a). Detection of obstructive sleep apnoea by ECG signals using deep learning architectures. *European Signal Processing Conference*, *2021-Janua*(September), 1382–1386. https://doi.org/10.23919/Eusipco47968.2020.9287360

Almutairi, H., Hassan, G. M., & Datta, A. (2021b). Detection of obstructive sleep apnoea by ECG signals using deep learning architectures. *European Signal Processing Conference*, *2021-Janua*, 1382–1386. https://doi.org/10.23919/Eusipco47968.2020.9287360

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. In *Journal of Big Data* (Vol. 8, Issue 1). Springer International Publishing. https://doi.org/10.1186/s40537-021-00444-8

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*(October 2019), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Benjafield, A. V, Eastwood, P. R., Heinzer, R., Morrell, M. J., Federal, U., Paulo, D. S., Paulo, S., & Valentine, K. (2019). Sleep Apnoea : a Literature-Based Analysis. *Lancet Respir Med*, *7*(8), 687–698. https://doi.org/10.1016/S2213-2600(19)30198-5.Estimation

Berry, R. B., Abreu, A. R., Krishnan, V., Quan, S. F., Strollo, P. J., & Malhotra, R. K. (2022). A transition to the American Academy of Sleep Medicine–recommended hypopnea definition in adults: initiatives of the Hypopnea Scoring Rule Task Force. *Journal of Clinical Sleep Medicine*, *18*(5), 1419–1425. https://doi.org/10.5664/jcsm.9952

Cen, L., Yu, Z. L., Kluge, T., & Ser, W. (2018). Automatic System for Obstructive Sleep Apnea Events Detection Using Convolutional Neural Network. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2018-July*, 3975–3978. https://doi.org/10.1109/EMBC.2018.8513363

Chan, E. D., Chan, M. M., & Chan, M. M. (2013). Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations. *Respiratory Medicine*, *107*(6), 789–799. https://doi.org/10.1016/j.rmed.2013.02.004

Chang, H. Y., Yeh, C. Y., Lee, C. Te, & Lin, C. C. (2020). A sleep apnea detection system based on a one-dimensional deep convolution neural network model using single-lead electrocardiogram. *Sensors (Switzerland)*, *20*(15), 1–15. https://doi.org/10.3390/s20154157

Chen, X., Wang, R., Zee, P., Lutsey, P. L., Javaheri, S., Alcántara, C., Jackson, C. L., Williams, M. A., & Redline, S. (2015). Racial/ethnic differences in sleep disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep*, *38*(6), 877–888. https://doi.org/10.5665/sleep.4732

Choi, S. H., Yoon, H., Kim, H. S., Kim, H. B., Kwon, H. Bin, Oh, S. M., Lee, Y. J., & Park, K. S. (2018a). Real-time apnea-hypopnea event detection during sleep by convolutional neural networks. *Computers in Biology and Medicine*, *100*, 123–131. https://doi.org/10.1016/j.compbiomed.2018.06.028

Choi, S. H., Yoon, H., Kim, H. S., Kim, H. B., Kwon, H. Bin, Oh, S. M., Lee, Y. J., & Park, K. S. (2018b). Real-time apnea-hypopnea event detection during sleep by convolutional neural networks. *Computers in Biology and Medicine*, *100*(June), 123–131. https://doi.org/10.1016/j.compbiomed.2018.06.028

Ciampiconi, L., Elwood, A., Leonardi, M., Mohamed, A., & Rozza, A. (2023). *A survey and taxonomy of loss functions in machine learning*. *1*(1), 1–29. http://arxiv.org/abs/2301.05579

Cohen, A. (2006). Biomedical signals: Origin and dynamic characteristics; frequency-domain analysis. In *Medical Devices and Systems*. https://doi.org/10.1201/9781420003864-6

Collop, N. A., Tracy, S. L., Kapur, V., Mehra, R., Kuhlmann, D., Fleishman, S. A., & Ojile, J. M. (2011). Obstructive sleep apnea devices for Out-Of-Center (OOC) testing: Technology evaluation. *Journal of Clinical Sleep Medicine*, *7*(5), 531–548. https://doi.org/10.5664/JCSM.1328

Cuesta, F. J. (2005). Nacional Sobre El Síndrome De Apneas-Hipopneas Del Sueño Grupo Español De Sueño (Ges). *Group*. http://www.sen.es/pdf/2005/consenso_sahs_completo.pdf

del Campo, F., Crespo, A., Cerezo-Hernández, A., Gutiérrez-Tobal, G. C., Hornero, R., & Álvarez, D. (2018). Oximetry use in obstructive sleep apnea. *Expert Review of Respiratory Medicine*, *12*(8), 665–681. https://doi.org/10.1080/17476348.2018.1495563

Dey, D., Chaudhuri, S., & Munshi, S. (2018). Obstructive sleep apnoea detection using convolutional neural network based deep learning framework. *Biomedical Engineering Letters*, *8*(1), 95–100. https://doi.org/10.1007/s13534-017-0055-y

eb16b8e82634249f8dd267a19caca25d3208c77a @ www.masimo.com. (n.d.). https://www.masimo.com/products/monitors/spot-check/mightysatrx/

Eguía, V. M., & Cascante, J. A. (2007). [Sleep apnea-hypopnea syndrome. Concept, diagnosis and medical treatment]. *Anales Del Sistema Sanitario de Navarra*, *30 Suppl 1*, 53–74. http://www.ncbi.nlm.nih.gov/pubmed/17486147

Eligulashvili, T. S., & Pal'man, A. D. (1997). Sleep apnea syndrome. *Klinicheskaia Meditsina*, *75*(9), 64–67.

Erdenebayar, U., Kim, Y. J., Park, J. U., Joo, E. Y., & Lee, K. J. (2019). Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram. *Computer Methods and Programs in Biomedicine*, *180*. https://doi.org/10.1016/j.cmpb.2019.105001

Ghamari, M. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics*, *4*(4). https://doi.org/10.15406/ijbsbe.2018.04.00125

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C., & Stanley, H. E. (2000). *Current Perspective*.

González Mangado, N., Egea-Santaolalla, C. J., Chiner Vives, E., & Mediano, O. (2020). Sleep Obstructive Apnea. *Open Respiratory Archives*, *2*(2), 46–66. https://doi.org/10.1016/j.opresp.2020.03.008

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. 1–17. http://arxiv.org/abs/2008.05756

Guilleminault, C., Eldridge, F. L., & Dement, W. C. (1973). Insomnia with sleep apnea: A new syndrome. *Science*, *181*(4102), 856–858. https://doi.org/10.1126/science.181.4102.856

Haidar, R., McCloskey, S., Koprinska, I., & Jeffries, B. (2018). Convolutional Neural Networks on Multiple Respiratory Channels to Detect Hypopnea and Obstructive Apnea Events. *Proceedings of the International Joint Conference on Neural Networks*, *2018-July*. https://doi.org/10.1109/IJCNN.2018.8489248

Ho, M. L., & Brass, S. D. (2011). Obstructive sleep apnea. *Neurology International*, *3*(3), 60–67. https://doi.org/10.4081/ni.2011.e15

Hu, S., Cai, W., Gao, T., & Wang, M. (2022). A Hybrid Transformer Model for Obstructive Sleep Apnea Detection Based on Self-Attention Mechanism Using Single-Lead ECG. *IEEE Transactions on Instrumentation and Measurement*, *71*. https://doi.org/10.1109/TIM.2022.3193169

Huttunen, R., Leppanen, T., Duce, B., Arnardottir, E. S., Nikkonen, S., Myllymaa, S., Toyras, J., & Korkalainen, H. (2023). A Comparison of Signal Combinations for Deep Learning-Based Simultaneous Sleep Staging and Respiratory Event Detection. *IEEE Transactions on Biomedical Engineering*, *70*(5), 1704–1714. https://doi.org/10.1109/TBME.2022.3225268

index @ www.nonin.com. (n.d.). https://www.nonin.com/

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, *33*(4), 917–963. https://doi.org/10.1007/s10618-019-00619-1

Kapoor, M., & Greenough, G. (2015). Home sleep tests for obstructive sleep apnea (OSA). *Journal of the American Board of Family Medicine*, *28*(4), 504–509. https://doi.org/10.3122/jabfm.2015.04.140266

Kapur, V. K., Auckley, D. H., Chowdhuri, S., Kuhlmann, D. C., Mehra, R., Ramar, K., & Harrod, C. G. (2017). Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: An American academy of sleep medicine clinical practice

guideline. *Journal of Clinical Sleep Medicine*, *13*(3), 479–504. https://doi.org/10.5664/jcsm.6506

Kim, P. (2017). MATLAB Deep Learning. In *MATLAB Deep Learning*. https://doi.org/10.1007/978-1-4842-2845-6

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krstinić, D., Braović, M., Šerić, L., & Božić-Štulić, D. (2020). *Multi-label Classifier Performance Evaluation with Confusion Matrix*. 01–14. https://doi.org/10.5121/csit.2020.100801

Kukačka, J., Golkov, V., & Cremers, D. (2017). *Regularization for Deep Learning: A Taxonomy*. http://arxiv.org/abs/1710.10686

Kulkas, A., Duce, B., Leppänen, T., Hukins, C., & Töyräs, J. (2017). Severity of desaturation events differs between hypopnea and obstructive apnea events and is modulated by their duration in obstructive sleep apnea. *Sleep and Breathing*, *21*(4), 829–835. https://doi.org/10.1007/s11325-017-1513-6

Kulkas, A., Tiihonen, P., Julkunen, P., Mervaala, E., & Töyräs, J. (2013). Desaturation delay, parameter for evaluating severity of sleep disordered breathing. *IFMBE Proceedings*, *39 IFMBE*, 336–339. https://doi.org/10.1007/978-3-642-29305-4_90

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leino, A., Nikkonen, S., Kainulainen, S., Korkalainen, H., Töyräs, J., Myllymaa, S., Leppänen, T., Ylä-Herttuala, S., Westeren-Punnonen, S., Muraja-Murro, A., Jäkälä, P., Mervaala, E., & Myllymaa, K. (2021). Neural network analysis of nocturnal SpO2 signal enables easy screening of sleep apnea in patients with acute cerebrovascular disease. *Sleep Medicine*, *79*, 71–78. https://doi.org/10.1016/j.sleep.2020.12.032

Liu, H., Cui, S., Zhao, X., & Cong, F. (2023). Detection of obstructive sleep apnea from single-channel ECG signals using a CNN-transformer architecture. *Biomedical Signal Processing and Control*, *82*(November 2022), 104581. https://doi.org/10.1016/j.bspc.2023.104581

Lloberes, P., Durán-Cantolla, J., Martínez-García, M. Á., Marín, J. M., Ferrer, A., Corral, J., Masa, J. F., Parra, O., Alonso-Álvarez, M. L., & Terán-Santos, J. (2011). Diagnóstico y tratamiento del síndrome de apneas-hipopneas del sueño. *Archivos de Bronconeumologia*, *47*(3), 143–156. https://doi.org/10.1016/j.arbres.2011.01.001

Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, *226*, 107161. https://doi.org/10.1016/j.cmpb.2022.107161

Madhan Mohan, P., Annie Nisha, A., Nagarajan, V., & Smiley Jeya Jothi, E. (2016). Measurement of arterial oxygen saturation (SpO2) using PPG optical sensor. *International Conference on Communication and Signal Processing, ICCSP 2016*, 1136–1140. https://doi.org/10.1109/ICCSP.2016.7754330

Mostafa, S. S., Mendonca, F., Ravelo-Garcia, A. G., Julia-Serda, G., & Morgado-Dias, F. (2020). Multi-Objective Hyperparameter Optimization of Convolutional Neural Network for Obstructive Sleep Apnea Detection. *IEEE Access*, *8*, 129586–129599. https://doi.org/10.1109/ACCESS.2020.3009149

Mostafa, S. S., Mendonça, F., Ravelo-García, A. G., & Morgado-Dias, F. (2019). A systematic review of detecting sleep apnea using deep learning. *Sensors (Switzerland)*, *19*(22), 1–26. https://doi.org/10.3390/s19224934

Mukherjee, D., Dhar, K., Schwenker, F., & Sarkar, R. (2021). Ensemble of deep learning models for sleep apnea detection: An experimental study. *Sensors*, *21*(16), 1–17. https://doi.org/10.3390/s21165425

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. https://doi.org/10.1186/s40537-014-0007-7

Nasifoglu, H., & Erogul, O. (2021). Obstructive sleep apnea prediction from electrocardiogram scalograms and spectrograms using convolutional neural networks. *Physiological Measurement*, *42*(6). https://doi.org/10.1088/1361-6579/ac0a9c

Netzer, N., Eliasson, A. H., Netzer, C., & Kristo, D. A. (2001). Overnight pulse oximetry for sleep-disordered breathing in adults: A review. *Chest*, *120*(2), 625–633. https://doi.org/10.1378/chest.120.2.625

Nikkonen, S., Korkalainen, H., Leino, A., Myllymaa, S., Duce, B., Leppanen, T., & Toyras, J. (2021). Automatic Respiratory Event Scoring in Obstructive Sleep Apnea Using a Long Short-Term Memory Neural Network. *IEEE Journal of Biomedical and Health Informatics*, *25*(8), 2917–2927. https://doi.org/10.1109/JBHI.2021.3064694

Nitzan, M., Romem, A., & Koppel, R. (2014). Pulse oximetry: Fundamentals and technology update. *Medical Devices: Evidence and Research*, *7*(1), 231–239. https://doi.org/10.2147/MDER.S47319

Otero, A., Félix, P., Barro, S., & Zamarrón, C. (2012). A structural knowledge-based proposal for the identification and characterization of apnoea episodes. *Applied Soft Computing Journal*, *12*(1), 516–526. https://doi.org/10.1016/j.asoc.2011.08.009

Pase, M. P., Harrison, S., Misialek, J. R., Kline, C. E., Cavuoto, M., & Baril, A. (2023). *Sleep Architecture , Obstructive Sleep Apnea , and Cognitive Function in Adults*. *6*(7), 1–14. https://doi.org/10.1001/jamanetworkopen.2023.25152

Pathinarupothi, R. K., Dhara Prathap, J., Rangan, E. S., Gopalakrishnan, A. E., Vinaykumar, R., & Soman, K. P. (2017). Single Sensor Techniques for Sleep Apnea Diagnosis Using Deep Learning. *Proceedings - 2017 IEEE International Conference on Healthcare Informatics, ICHI 2017*, 524–529. https://doi.org/10.1109/ICHI.2017.37

Penzel, T. (2020). *Advances in the Diagnosis and Treatment of Sleep Apnea*.

Penzel, T., Moody, G. B., Mark, R. G., Goldberger, A. L., & Peter, J. H. (2000). Apnea-ECG database. *Computers in Cardiology*, 255–258.

Qin, H., & Liu, G. (2022). A dual-model deep learning method for sleep apnea detection based on representation learning and temporal dependence. *Neurocomputing*, *473*, 24–36. https://doi.org/10.1016/j.neucom.2021.12.001

Rasamoelina, A. D., Adjailia, F., & Sincak, P. (2020). A Review of Activation Function for Artificial Neural Network. *SAMI 2020 - IEEE 18th World Symposium on Applied Machine Intelligence and Informatics, Proceedings*, 281–286. https://doi.org/10.1109/SAMI48414.2020.9108717

Ruder, S. (2016). *An overview of gradient descent optimization algorithms*. 1–14. http://arxiv.org/abs/1609.04747

Rundo, J. V., & Downey, R. (2019). Polysomnography. *Handbook of Clinical Neurology*, *160*(1877), 381–392. https://doi.org/10.1016/B978-0-444-64032-1.00025-4

Saxena, A. (2022). An Introduction to Convolutional Neural Networks. *International Journal for Research in Applied Science and Engineering Technology*, *10*(12), 943–947. https://doi.org/10.22214/ijraset.2022.47789

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Sharan, R. V., Berkovsky, S., Xiong, H., & Coiera, E. (2020). ECG-Derived Heart Rate Variability Interpolation and 1-D Convolutional Neural Networks for Detecting Sleep Apnea. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, *2020-July*, 637–640. https://doi.org/10.1109/EMBC44109.2020.9175998

Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697857

Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. (2023). *A Comprehensive Overview and Comparative Analysis on Deep Learning Models: CNN, RNN, LSTM, GRU*. http://arxiv.org/abs/2305.17473

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability.1. Shrout PE, Fleiss JL: Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979, 86:420–8. *Psychological Bulletin*, *86*(2), 420–428. http://www.ncbi.nlm.nih.gov/pubmed/18839484

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Urtnasan, E., Park, J. U., & Lee, K. J. (2018). Multiclass classification of obstructive sleep apnea/hypopnea based on a convolutional neural network from a single-lead electrocardiogram. *Physiological Measurement*, *39*(6). https://doi.org/10.1088/1361-6579/aac7b7

Vaquerizo-Villar, F., Alvarez, D., Kheirandish-Gozal, L., Gutierrez-Tobal, G. C., Barroso-Garcia, V., Santamaria-Vazquez, E., Campo, F. Del, Gozal, D., & Hornero,

R. (2021). A Convolutional Neural Network Architecture to Enhance Oximetry Ability to Diagnose Pediatric Obstructive Sleep Apnea. *IEEE Journal of Biomedical and Health Informatics*, *25*(8), 2906–2916. https://doi.org/10.1109/JBHI.2020.3048901

Wang, T., Lu, C., Shen, G., & Hong, F. (2019). Sleep apnea detection from a single-lead ECG signal with automatic feature-extraction through a modified LeNet-5 convolutional neural network. *PeerJ*, *2019*(9), 1–17. https://doi.org/10.7717/peerj.7731

Wang, X., Cheng, M., Wang, Y., Liu, S., Tian, Z., Jiang, F., & Zhang, H. (2020). Obstructive sleep apnea detection using ecg-sensor with convolutional neural networks. *Multimedia Tools and Applications*, *79*(23–24), 15813–15827. https://doi.org/10.1007/s11042-018-6161-8

Ye, J. C. (2022). Convolutional Neural Networks. In *Mathematics in Industry* (Vol. 37). https://doi.org/10.1007/978-981-16-6046-7_7

Zhang, G. Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., Mariani, S., Mobley, D., & Redline, S. (2018). The National Sleep Research Resource: Towards a sleep data commons. *Journal of the American Medical Informatics Association*, *25*(10), 1351–1358. https://doi.org/10.1093/jamia/ocy064

Zhang, J., Tang, Z., Gao, J., Lin, L., Liu, Z., Wu, H., Liu, F., & Yao, R. (2021). Automatic detection of obstructive sleep apnea events using a deep CNN-LSTM model. *Computational Intelligence and Neuroscience*, *2021*. https://doi.org/10.1155/2021/55947

.

GLOSARIO DE SIGLAS Y ACRÓNIMOS

A: Apnea
AASM: *American Academy of Sleep Medicine*
AC: Componente Alterno
ACC: *Accuracy*
Adam: *Adaptative Moment Estimation*
AED: Apnea-ECG
AHÍ : *Apnea-Hypopnea Index*
ANN: *Artificial Neural Network*
AOS: Apnea Obstructiva del Sueño
AUC: *Area Under Curve*
BiGRU: *Bidirectional Gated Recurrent Unit*
CAM: *Class Activation Mapping*
CNN: *Convolutional Neural Network*
DC: Componente continuo
DL: *Deep Learning*
DNN: *Deep Neural Network*
ECG: Electrocardiograma
ECV: Enfermedades Cardiovasculares
EEG: Electroencefalograma
EMG: Electromiograma
EOG: Electrooculograma
e/h: Eventos por hora
FA: Flujo aéreo
FC: *Fully-Connected*
FFN: *Feed-Fodward Network*
FN: Falso negativo
FP: Falso positivo
GAP: *Global Average Pooling*
Grad-CAM: *Gradient-weighted Class Activation Mapping*
GRU: *Gated Recurrent Unit*
H: Hipopnea
HbO_2: Hemoglobina oxigenada
HHb: Hemoglobina desoxigenada
HSAT: *Home Sleep Apnea Testing*
IA: Inteligencia Artificial
Kappa: Coeficiente kappa de Cohen
LR: *Learning Rate*
LSTM: *Long-Short Term Memory*
MESA: *Multi-Ethnic Study of Atherosclerosis*
ML: Machine Learning
MSC: *Mean Square Colums*
MSE: *Mean Square Error*
MSR: *Mean Square Subject*
N: Normal
NPV: *Negative Predictive Value*
OSA: *Obstructive Sleep Apnea*
O_2: Oxígeno
PAP: Presión positiva en las vías respiratorias
PO: Pulsioximetría
PPG: Fotopletismografía
PPV: *Positive Predictive Value*
PR: *Pulse rate*
PSG: Polisomnografía
REI: *Respiratory Event Index*
ReLU: *Rectified Linear Unit*
RNN: *Recurrent Neural Network*
SaO_2: Saturación de oxígeno en sangre
Se: Sensibilidad
Sp: Especificidad
SpO_2: Saturación periférica de oxígeno en sangre
Tanh: Tangente hiperbólica
TFG: Trabajo de Fin de Grado
TRS: Trastorno Respiratorio del Sueño
UCD: *College Dublin Sleep Apnea Databse*
VAS: Vías Respiratorias Superiores
VN: Verdadero negativo
VP: Verdadero positivo
XAI: *Explainable Artificial Intelligence*