UNIVERSIDAD DE VALLADOLID

FACULTAD DE MEDICINA

ESCUELA DE INGENIERÍAS INDUSTRIALES

TRABAJO DE FIN DE GRADO

GRADO EN INGENIERÍA BIOMÉDICA

Deep Learning explicable para el diagnóstico de la apnea del sueño en niños

Autora:

Dª. María Ruiz Redondo

Tutores:

D. Gonzalo César Gutiérrez Tobal

Dª. María Herrero Tudela

Valladolid, 11 de septiembre de 2025

TÍTULO:	Deep Learning explicable para el diagnóstico de la apnea del sueño en niños
AUTORA:	**D.ª María Ruiz Redondo**
TUTORES:	**D. Gonzalo César Gutiérrez Tobal** **D.ª María Herrero Tudela**
DEPARTAMENTO:	**Teoría de la Señal y Comunicaciones e Ingeniería Telemática**

TRIBUNAL

PRESIDENTE:	Dr. D. Gonzalo César Gutiérrez Tobal
SECRETARIO:	Dra. Dª. María García Gadañón
VOCAL:	Dr. D. Daniel Álvarez González
SUPLENTE 1:	Dr. D. Javier Gómez Pilar
SUPLENTE 2:	Dr. D. Carlos Gómez Peña

AGRADECIMIENTOS

En primer lugar, me gustaría agradecer a mis tutores Gonzalo César Gutiérrez Tobal y María Herrero Tudela por guiarme durante esta etapa académica. Su dedicación ha sido fundamental para ayudarme a superar las dificultades del trabajo. Gracias por haberme mostrado esta área tan interesante de la Ingeniería Biomédica.

También me gustaría dar las gracias al Grupo de Ingeniería Biomédica de la Universidad de Valladolid. En especial me gustaría agradecer a mis compañeros de laboratorio, que han estado a mi lado ofreciéndome su ayuda y compañía, contribuyendo tanto en lo profesional como en lo personal.

Por último, deseo agradecer a toda mi familia y amigos, por acompañarme, apoyarme y escucharme con tanta paciencia. Especialmente a mi hermana que me ha dado todos los consejos necesarios para poder completar este trabajo.

RESUMEN

La apnea obstructiva del sueño (AOS) pediátrica es un trastorno respiratorio nocturno que afecta significativamente a la salud y calidad de vida de los niños, incluyendo su desarrollo cognitivo. El diagnóstico de referencia es la polisomnografía (PSG), pero presenta limitaciones, como una disponibilidad reducida, ser de alto coste y resultar incómoda para los pacientes. Ante estas dificultades, existe una amplia variedad de métodos de *deep learning* que emplean la señal de oximetría (SpO_2) para el diagnóstico automático de la AOS infantil. Sin embargo, el número de estudios en los que se usa imagen procedente de la señal de SpO_2 y métodos de *eXplainable Artificial Intelligence* (XAI) es escaso. En este trabajo se propone un nuevo método de *deep learning* explicable para el diagnóstico y estadificación de la AOS en niños a partir de espectrogramas generados de señales de SpO_2. Para ello, se ha utilizado una base de datos semipública de 1.638 registros del sueño procedentes del estudio *Childhood Adenotonsillectomy Trial* (CHAT). El enfoque *deep learning* se ha llevado a cabo mediante *transfer learning* de la arquitectura ResNet50 de dos formas: ResNet50 con pre-entreno con imágenes de *Imagenet* y ResNet50 con pre-entreno con imágenes de retina, ambas para estimar el índice apnea-hipopnea (AHI). Además, se ha utilizado el método *Shapley Additive exPlanation*s (SHAP) para entender en qué se han basado los modelos evaluados para la toma de decisiones. Los resultados del mejor modelo, ResNet50 con pre-entreno con *Imagenet*, presentan una precisión de cuatro clases del 68.30% y un coeficiente kappa de 0.5326. La sensibilidad y especificidad obtenidas para cada umbral de severidad del AHI son, respectivamente, 84.10% y 71.64% para el umbral 1 e/h, 78.02% y 95.81% para el umbral de 5 e/h y 78.57% y 97.73% para el umbral de 10 e/h. El análisis SHAP explica cómo el modelo otorga importancia a zonas de aumento de potencia en zonas de bajas frecuencias debido a eventos apneicos asociados. Además, también muestra la razón detrás de algunos de los errores cometidos, al dar importancia a desaturaciones no ligadas a marcas de apnea e hipopnea, fluctuaciones de la señal o al ignorar eventos cercanos a artefactos en la señal. Los altos resultados alcanzados y la utilización de métodos de XAI proporcionan mayor transparencia y confianza en las estimaciones de los modelos. El empleo de este método de diagnóstico con imagen procedente de señal de SpO_2 se presenta como una herramienta accesible e innovadora para el diagnóstico de la AOS en niños como alternativa a la PSG.

Palabras clave apnea obstructiva del sueño (AOS), saturación de oxígeno en sangre (SpO_2), oximetría, niños, imagen, espectrograma, *deep learning*, redes neuronales convolucionales (CNN), *explainable artificial intelligence* (XAI).

ABSTRACT

Obstructive sleep apnea (OSA) in children is a respiratory disorder that significantly affects patients' health and quality of life, including cognitive impairment. The reference diagnosis is polysomnography (PSG), but it has limitations such as restricted availability, high cost and inconvenience for patients. Given these difficulties, there is a wide variety of deep learning methods use oximetry signals (SpO_2) for the automatic diagnosis of pediatric OSA. However, the number of studies that use images derived from SpO_2 and eXplainable Artificial Intelligence (XAI) methods is scarce. This work proposes a new explainable deep learning method for diagnosing and staging OSA in children based on spectrograms generated from SpO_2 signals. Accordingly, 1.638 sleep recordings have been used from the semi-public Childhood Adenotonsillectomy Trial (CHAT) database. Transfer learning with ResNet50 has been carried out in two ways: ResNet50 with pre-training with Imagenet images and ResNet50 with pre-training with retinal images, both to estimate the apnea-hypopnea index (AHI). In addition, Shapley Additive exPlanations (SHAP) has been used to explain and understand the model's decisions. The results of the best model, ResNet50 with pre-training with Imagenet, show a four-class accuracy of 68.30% and a kappa index of 0.5326. The sensitivity and specificity obtained for each AHI severity threshold are, respectively, 84.10% and 71.64% for 1 e/h, 78.02% and 95.81% for the 5 e/h threshold, and 78.57% and 97.73% for the 10 e/h threshold. SHAP explains how the model assigns importance to areas of increased power in low-frequency zones due to the apneic events. Furthermore, it also shows the reasons behind some of the errors made, by giving importance to desaturations not related to apnea and hypopnea events, signal fluctuations, or by ignoring events close to artifacts in the signal. The high results reached and the use of XAI methods provide greater transparency and confidence in the model estimates. The use of this image diagnostics method derived from SpO_2 signal presents itself as an accessible and innovative tool for the diagnosis of OSA in children as an alternative to PSG.

Keywords obstructive sleep apnea (OSA), blood oxygen saturation (SpO_2), oximetry, children, image, spectrogram, deep learning, convolutional neuronal networks (CNN), explainable artificial intelligence (XAI).

ÍNDICE GENERAL

ÍNDICE GENERAL I
ÍNDICE DE FIGURAS III
ÍNDICE DE TABLAS V
CAPÍTULO 1: INTRODUCCIÓN 1
1.1 APNEA OBSTRUCTIVA DEL SUEÑO 1
1.2 DIAGNÓSTICO 2
1.3 OXIMETRÍA Y SEÑAL DE SATURACIÓN DE OXÍGENO 4
1.4 ESTADO DEL ARTE DEL DIAGNÓSTICO AUTOMÁTICO DE LA AOS A PARTIR DE IMAGEN 6
1.4.1 DEEP LEARNING 8
1.4.2 EXPLAINABLE ARTIFICIAL INTELLIGENCE 9
1.5 COMPARACIÓN Y ELECCIÓN DEL MÉTODO A IMPLEMENTAR 10
1.6 HIPÓTESIS Y OBJETIVOS 11
1.7 PLAN DE TRABAJO Y ESTRUCTURA DEL TFG 11
1.7.1 PLAN DE TRABAJO 11
1.7.2 ESTRUCTURA DEL TFG 12
CAPÍTULO 2: SUJETOS E IMÁGENES DEL ESTUDIO 15
2.1 BASE DE DATOS 15
2.2 PARTICIÓN DEL CONJUNTO DE DATOS 15
CAPÍTULO 3: METODOLOGÍA 19
3.1 TRANSFORMACIÓN DE LA SEÑAL EN IMAGEN 19
3.2 REDES NEURONALES CONVOLUCIONALES 20
3.2.1 RESNET 22
3.3 SHAPLEY ADDITIVE EXPLANATIONS 23
3.4 IMPLEMENTACIÓN DE LOS MÉTODOS 24
3.4.1 TRANSFER LEARNING CON RESNET50 24
3.4.2 TRANSFER LEARNING CON RESNET PRENTRENADA CON IMÁGENES DE RETINA 26
3.4.3 IMPLEMENTACIÓN DE SHAP 27
3.6 ANÁLISIS ESTADÍSTICO 28

3.5.1 MÉTRICAS DE RENDIMIENTO DE LA ESTIMACIÓN DEL AHI 28
3.5.2 MÉTRICAS DE RENDIMIENTO DE LA CLASIFICACIÓN 29
CAPÍTULO 4: RESULTADOS 33
4.1 OPTIMIZACIÓN DE HIPERPARÁMETROS 33
4.2 RESULTADOS DE LOS MEJORES MODELOS 34
4.3 RESULTADOS DE SHAP 37
CAPÍTULO 5: DISCUSIÓN DE LOS RESULTADOS 47
5.1 CLASIFICACIÓN DE LA SEVERIDAD DE LA APNEA 47
5.2 ESTIMACIÓN DEL AHI 48
5.3 INTERPRETACIÓN MEDIANTE SHAP 48
5.3.1 CLASE NO-AOS 49
5.3.2 CLASE AOS LEVE 52
5.3.3 CLASE AOS MODERADA 55
5.3.4 CLASE AOS SEVERA 57
5.4 COMPARACIÓN CON OTROS ESTUDIOS 61
5.5 LIMITACIONES 66
CAPÍTULO 6: CONCLUSIONES Y LINEAS FUTURAS 67
6.1 CONTRIBUCIONES 67
6.2 CONCLUSIONES 68
6.3 LINEAS FUTURAS 69
REFERENCIAS 71

ÍNDICE DE FIGURAS

Figura 1. Señal SpO_2 de la base de datos CHAT. 5
Figura 2. Esquema del proceso de transformación de la señal a imagen y separación de los datos. 17
Figura 3. Espectrograma de señal SpO_2 con mapa de color Hot y ventana de 2 minutos. 20
Figura 4. Ejemplo de una operación de convolución con kernel de 3x3 sobre una imagen de 5x5 píxeles. 21
Figura 5. Esquema de un bloque residual de Resnet (Xu, Fu and Zhu, 2023) 23
Figura 6. Esquema del procedimiento empleado en el estudio desde la señal hasta la estadificación de la enfermedad. 26
Figura 7. Matriz de confusión del grupo de test obtenida del mejor modelo Resnet50 entrenada con espectrogramas de la señal SpO_2 de la base de datos CHAT. 34
Figura 8. Diagrama Bland-Altman para modelo el mejor modelo ResNet50. 35
Figura 9. Matriz de confusión obtenida del mejor modelo Resnet50 + Retina entrenada con espectrogramas de la señal SpO_2 de la base de datos CHAT. 36
Figura 10.Diagrama Bland-Altman para modelo el mejor modelo ResNet50 + Retina. 37
Figura 11. Análisis SHAP para el sujeto de clase No-AOS con la mejor estimación del AHI. 39
Figura 12. Análisis SHAP para el sujeto de clase No-AOS con la peor estimación del AHI. 40
Figura 13. Análisis SHAP para el sujeto de clase AOS leve con la mejor estimación del AHI. 41
Figura 14. Análisis SHAP para el sujeto de clase AOS leve con la peor estimación del AHI. 42
Figura 15. Análisis SHAP para el sujeto de clase AOS moderada con la mejor estimación del AHI. 43
Figura 16. Análisis SHAP para el sujeto de clase AOS moderada con la peor estimación del AHI. 44
Figura 17. Análisis SHAP para el sujeto de clase AOS severa con la mejor estimación del AHI. 45
Figura 18. Análisis SHAP para el sujeto de clase AOS severa con la peor estimación del AHI. 46
Figura 19. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase No-AOS con la mejor estimación del AHÍ. 49
Figura 20. Zoom realizado en una zona del espectrograma de la mejor estimación para No-AOS con mapa SHAP por encima. 50
Figura 21.Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase No-AOS con la peor estimación del AHI. 51

Figura 22.Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase No-AOS con la peor estimación del AHI. 51
Figura 23. Zoom realizado en una zona del espectrograma de la peor estimación para No-AOS con mapa SHAP por encima. 52
Figura 24. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS leve con la mejor estimación del AHI. 53
Figura 25. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS leve con la mejor estimación del AHI. 53
Figura 26. Zoom realizado en una zona del espectrograma de la mejor estimación para AOS leve con mapa SHAP por encima. 54
Figura 27. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS leve con la peor estimación del AHI. 54
Figura 28. Zoom realizado en una zona del espectrograma de la peor estimación para AOS leve con mapa SHAP por encima. 55
Figura 29. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS moderada con la mejor estimación del AHI. 55
Figura 30. Zoom realizado en una zona del espectrograma de la mejor estimación para AOS moderada con mapa SHAP por encima 56
Figura 31. Zoom realizado en una zona del espectrograma en color de la mejor estimación para AOS moderada. 56
Figura 32. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS moderada con la peor estimación del AHI. 57
Figura 33. Zoom realizado en una zona del espectrograma de la peor estimación para AOS moderada con mapa SHAP por encima 57
Figura 34. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS severa con la mejor estimación del AHI. 58
Figura 35. Zoom realizado en una zona del espectrograma de la mejor estimación para AOS severa con mapa SHAP por encima. 58
Figura 36. Zoom realizado en una zona del espectrograma en color de la mejor estimación para AOS severa. 59
Figura 37. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS severa con la peor estimación del AHI. 59
Figura 38. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS severa con la peor estimación del AHI. 60
Figura 40. Zoom realizado en una zona del espectrograma en color de la peor estimación para AOS. 60
Figura 40. Zoom realizado en una zona del espectrograma en color de la peor estimación para AOS. 60

ÍNDICE DE TABLAS

Tabla 1. Estado del arte del empleo de imagen junto técnicas de DL y ML para el diagnóstico de la AOS. 8
Tabla 2. Tabla con datos sociodemográficos de la base de datos CHAT 17
Tabla 3. Coeficiente kappa para cada combinación del número de neuronas de las capas Dense en ResNet50. 33
Tabla 4. Coeficiente kappa para cada combinación del número neuronas de las capas Dense en ResNet50 pre-entrenada con imágenes de retina. 33
Tabla 5. Métricas calculadas para el mejor modelo ResNet50para cada umbral. 35
Tabla 6. ICC para el mejor modelo ResNet50. 35
Tabla 7. Métricas calculadas para el mejor modelo ResNet50 + Retina para cada umbral. 36
Tabla 8. ICC para el mejor modelo ResNet50 + Retina. 36
Tabla 9. Mejores y peores estimaciones del AHI para cada clase de apnea. 37
Tabla 10. Comparación de los resultados del modelo ResNet50 con otros estudios..... 65

CAPÍTULO 1: INTRODUCCIÓN

1.1 APNEA OBSTRUCTIVA DEL SUEÑO

La apnea obstructiva del sueño (AOS) es un trastorno respiratorio que afecta significativamente a la salud y al desarrollo cognitivo de los niños (Marcus *et al.*, 2012). Aunque depende del método diagnóstico y la edad estudiada, su prevalencia se estima entre el 1 y 5 % de la población pediátrica y se presenta con mayor frecuencia en niños con obesidad e hipertrofia adenoamigdalar (Marcus *et al.*, 2012). Tanto el diagnóstico de esta enfermedad como el tratamiento temprano es importante para la mejora del rendimiento académico, social y la calidad de vida de los pacientes (Marcus *et al.*, 2012).

La AOS se caracteriza por episodios repetitivos de obstrucción parcial (hipopnea) o total (apnea) de la vía aérea superior (VAS) durante el sueño y su origen está vinculado a factores tanto anatómicos como funcionales (C. L. Marcus et al., 2012; Tauman & Gozal, 2011). El estrechamiento o colapso de la vía aérea provoca alteraciones en la oxigenación, por lo que puede generar una serie de secuelas como problemas neuroconductuales, disfunciones metabólicas, efectos cardiovasculares, inflamación sistémica o dificultades en el crecimiento (C. L. Marcus et al., 2012; Tauman & Gozal, 2011).

Esta enfermedad se relaciona con diferentes síntomas entre los que destaca la llamada *triada principal* constituida por ronquidos fuertes y frecuentes, pausas en la respiración y somnolencia diurna (Lloberes *et al.*, 2011). Además, los niños con AOS pueden presentar dificultad para respirar durante el sueño, y, junto con los síntomas anteriores, lleva en ocasiones a posturas anormales, respiración bucal y una sudoración excesiva. (Liukkonen *et al.*, 2012). Otros signos frecuentes incluyen movimientos anómalos de tórax, despertares frecuentes y sueño intranquilo (Tauman and Gozal, 2011). La frecuencia de algunos síntomas depende de la edad, por ejemplo, la somnolencia diurna excesiva es el síntoma más común en los pacientes adolescentes mientras que en niños más pequeños son la hiperactividad y los problemas de atención (Chang and Chae, 2010).

El tratamiento de la AOS varía en función de cada paciente y de la gravedad de la enfermedad. En niños, el tratamiento preferente es la adenoamigdalectomía, ya que muchos de los casos se deben a la hipertrofia adenoamigdalar, y este procedimiento consiste en la extirpación de amígdalas y vegetaciones. Es un tratamiento con pocas complicaciones, pero, en ocasiones, presenta contraindicaciones o hay escasa respuesta y la AOS es persistente (Lloberes et al., 2011; C. L. Marcus et al., 2012). En estos casos, se opta por el tratamiento con presión positiva continua en la vía respiratoria (*Continuous Positive Airway Pressure*, CPAP), que suele ser la primera opción en adultos debido a su alta efectividad (Lloberes et al., 2011; C. L. Marcus et al., 2012). El tratamiento de elección siempre se acompaña de un seguimiento y de medidas higiénico-dietéticas, positivas particularmente en los casos de obesidad (Lloberes *et al.*, 2011).

Como se ha indicado anteriormente, la hipertrofia de las adenoides y amígdalas es el principal factor de riesgo en población infantil, ya que el exceso de tejido en estas estructuras genera un estrechamiento de la VAS durante el sueño (Gulotta *et al.*, 2019). Además, existen otros factores que pueden aumentar la probabilidad de padecer AOS, como la obesidad, anomalías craneofaciales, rinitis alérgica, prematuridad, factores genéticos, factores raciales o el sexo (Gulotta *et al.*, 2019).

1.2 DIAGNÓSTICO

El diagnóstico de la AOS en población infantil se basa en los resultados de las pruebas objetivas combinadas con la evaluación clínica. La evaluación clínica consiste en la identificación de los síntomas presentados, la exploración física y los cuestionarios del sueño como el *Pediatric Sleep Questonnaire* (PSQ). Sin embargo, este conjunto de información no da respuestas lo suficientemente precisas como para establecer un diagnóstico final y se debe apoyar en los resultados de las pruebas diagnósticas objetivas, como la polisomnografía (PSG) nocturna (Chang and Chae, 2010; Gulotta *et al.*, 2019).

La PSG se considera el *gold standard* diagnóstico ya que es capaz de recoger tanto el número de eventos de apnea e hipopnea como sus efectos reflejados en diferentes señales fisiológicas (Lloberes *et al.*, 2011; Marcus *et al.*, 2012). Entre estas señales se encuentra el flujo aéreo, movimientos toracoabdominales, electrocardiograma (ECG), electrooculograma (EOG), electroencefalograma (EEG) o electromiograma (EMG) (Chang and Chae, 2010; Lloberes *et al.*, 2011). Además, con este método, un solo registro suele ser suficiente para un diagnóstico eficaz y la posterior evaluación terapéutica. Sin embargo, la PSG también presenta inconvenientes como tener una disponibilidad limitada, ser de alto coste y poder provocar incomodidad a los pacientes, especialmente en los niños (Lloberes *et al.*, 2011).

Por esta razón, se buscan otras herramientas de diagnóstico alternativas que contrarresten estas desventajas de forma eficiente. Un enfoque habitual ha sido la evaluación en solitario de las señales biomédicas procedentes de la PSG para realizar diagnósticos automáticos, aunque todavía se presenta como un campo en desarrollo, y la mayoría de los estudios se centran en adultos. El uso de imágenes derivada de señales biomédicas se ha basado principalmente en la transformación del ECG dado que esta permite detectar variaciones de la frecuencia cardiaca, arritmias y otras disfunciones en relación con los episodios de apnea (Mashrur *et al.*, 2021; Nasifoglu and Erogul, 2021; Niroshana *et al.*, 2021; Yu *et al.*, 2021; Gupta, Bajaj and Ansari, 2022; Ullah *et al.*, 2023; Lin *et al.*, 2024; Linh *et al.*, 2024; Zhou and Kang, 2024; Akter *et al.*, 2025; Choudhury *et al.*, 2025).

Otro tipo de señal es la saturación de oxígeno en sangre (SpO_2), que muestra las caídas de oxigenación durante el sueño debido a las pausas respiratorias (Mortazavi *et al.*, 2023; Stewart *et al.*, 2023). También se hace uso de la presión de aire nasal (NAP) que muestra los cambios de la respiración y el lugar anatómico de la obstrucción (Crowson *et al.*, 2023). Otra señal de estudio es el flujo de aire nasal (NA) para medir sus interrupciones,

tanto por sí solo como combinado con señales torácicas y abdominales (Wu *et al.*, 2021; Wang, Koprinska and Jeffries, 2023).

Además, se puede analizar el EMG, que recoge el tono muscular durante el sueño y detecta episodios de relajación excesiva asociados a eventos apneicos (Moradhasel *et al.*, 2023). El EEG registra la actividad cerebral y sus patrones, por lo que también es indicada para detectar desincronización, microdespertares o alteración de las etapas de sueño provocado por la AOS (Tanci and Hekim, 2025). Por último, los sonidos de la respiración también pueden resultar útiles para buscar la ausencia de respiraciones, atragantamientos o ronquidos (Romero *et al.*, 2022; Wang *et al.*, 2022; Song *et al.*, 2023).

En particular, entre todas estas señales, la SpO_2 es una señal que se emplea como cribado debido a que está involucrada en la definición de los eventos apnéicos. Además, aunque en la práctica no sustituye a la PSG, su accesibilidad y su fácil uso la convierte un método interesante para la detección de la AOS infantil (Kaditis, Kheirandish-Gozal and Gozal, 2015). Existen pocos estudios que utilicen imágenes derivadas de la señal de SpO_2 en población infantil. Sin embargo, la amplia disponibilidad de modelos de *deep learning* preentrenados en imágenes convierte a estas técnicas en una alternativa novedosa y prometedora para el diagnóstico de AOS en niños

Independientemente de cuál sea la herramienta seleccionada, para contar con un diagnóstico fiable de la AOS es frecuente atender a unos criterios fijos y estandarizados. La Academia Americana de Medicina del sueño (*American Academy of Sleep Medicine*, AASM) clasifica un evento como apnea obstructiva en niños si se da una caída del flujo de aire superior o igual al 90% respecto al valor base durante 2 respiraciones y está asociada con un esfuerzo respiratorio continuo o aumentado. En el caso de un evento de hipopnea en niños, se debe cumplir que la caída del flujo sea un 30% respecto al valor base durante el tiempo correspondiente a 2 respiraciones y estar asociada a una desaturación de oxígeno de 3% o un microdespertar (Berry *et al.*, 2018).

Así, el parámetro más empleado en el diagnóstico de AOS es el índice de apnea-hipopnea (*Apnea-Hypopnea Index*, AHI) que indica el número de apneas e hipopneas por hora de sueño y su cálculo según la AASM es el siguiente (Ecuación 1) (Berry *et al.*, 2018):

$$AHI = \frac{(Nº Apneas + Nº Hipopneas)\,.\,60}{Tiempo\ Total\ de\ Sueño\ (TST)(\text{min})} \tag{1}$$

La estadificación de la severidad de la apnea en niños se realiza utilizando el AHI. Los umbrales comúnmente empleados son (Gutiérrez-Tobal *et al.*, 2022):

- Sin AOS: $AHI \leq 1$
- AOS leve: $1 \leq AHI < 5$
- AOS moderada: $5 \leq AHI < 10$
- AOS grave: $AHI \geq 10$

1.3 OXIMETRÍA Y SEÑAL DE SATURACIÓN DE OXÍGENO

La oximetría es una forma diagnóstica no invasiva que recoge la señal de SpO_2 (Nitzan, Romem and Koppel, 2014). Esta señal se basa en la absorbancia de luz a través de los tejidos, obteniendo así una estimación del porcentaje de hemoglobina oxigenada en el cuerpo (Nitzan, Romem and Koppel, 2014). Esto se consigue gracias a un detector colocado normalmente en la palmas o dedos de las manos o pies, en el lóbulo de la oreja o en la frente, en función de la edad de los niños (Nitzan, Romem and Koppel, 2014; Hess, 2016; Al-Beltagi *et al.*, 2024).

Durante un evento de apnea o hipopnea, la obstrucción de las VAS puede reducir el oxígeno en sangre, lo que lleva a una disminución en el valor de la SpO_2. Estas desaturaciones intermitentes durante el sueño pueden ser registradas mediante oximetría nocturna a partir de distintos indicadores (Kaditis, Kheirandish-Gozal and Gozal, 2015).

Es posible calcular el índice de desaturación de oxígeno (*Oxygen Desaturation Index*, ODI), parámetro que representa el número de veces por hora que la saturación desciende un determinado porcentaje. Los más comunes son ODI3 y ODI4 para descensos del 3% y 4%, respectivamente. Asimismo, se pueden recoger clústeres, agrupaciones de estas desaturaciones, que ocurren de manera consecutiva durante 10-30 minutos (Kaditis, Kheirandish-Gozal and Gozal, 2015).

A partir de la señal de SpO_2 se distingue de manera indirecta los efectos de la AOS. Los valores que se consideran fuera de lo normal en niños son saturaciones por debajo del 95% y ODI4 $\geq$ 2.2 eventos/hora (e/h) (Kaditis, Kheirandish-Gozal and Gozal, 2015). Atendiendo a estos valores se pueden observar patrones capaces de detectar la presencia o ausencia de apnea y de cierta forma también guiar el tratamiento y predecir la respuesta a este (Kaditis, Kheirandish-Gozal and Gozal, 2015).

Aunque no siempre los eventos de apnea provoquen caídas en la saturación y el AHI se recoja mediante PSG, la información recogida por la señal de SpO_2 nos permite dar una estimación (Magalang et al., 2003). Varios indicadores de esta señal se encuentran correlacionados con el AHI y, aunque la oximetría pueda subestimar este índice, se presenta como una solución eficaz, especialmente en los casos más severos *(Magalang et al., 2003)*. A medida que el AHI se incrementa, también lo hace la probabilidad de que ocurran episodios de desaturación, así como la tendencia a que los niveles de saturación de oxígeno alcancen valores más bajos durante intervalos de tiempo más prolongados (Magalang *et al.*, 2003; Kaditis, Kheirandish-Gozal and Gozal, 2015; Wali *et al.*, 2020).

Dado que la señal de SpO_2 no constituye una medición directa, como lo es la saturación arterial de oxígeno (SaO_2), cuenta con un error máximo del 4%, pudiendo aumentar para valores de saturación por debajo del 70-80%. Esto se debe a que las calibraciones se realizan con pacientes sanos, y, por ello, en situaciones críticas, se recomienda usar otras pruebas adicionales (Nitzan, Romem and Koppel, 2014; Hess, 2016). Sin embargo, a diferencia de la SpO_2, la medición de SaO_2 es una técnica invasiva y se realiza de forma

más puntual. Una ventaja de la SpO_2 es que se trata de una prueba cómoda que se emplea a modo de monitorización y de forma continuada, siendo fiable en pacientes estables (Nitzan, Romem and Koppel, 2014; Hess, 2016).

En la Figura 1 se muestra un ejemplo de señal de SpO_2 a lo largo del tiempo. Se observa que los valores normales de oxigenación oscilan entre el 95% y el 100%, mientras que los descensos hacia valores más bajos, que se manifiestan en forma de picos, corresponden a episodios de desaturación. Cabe señalar que las caídas abruptas hasta el 0% son artefactos, componentes que no pertenecen a la señal, que no aportan información y que se deben a factores externos.

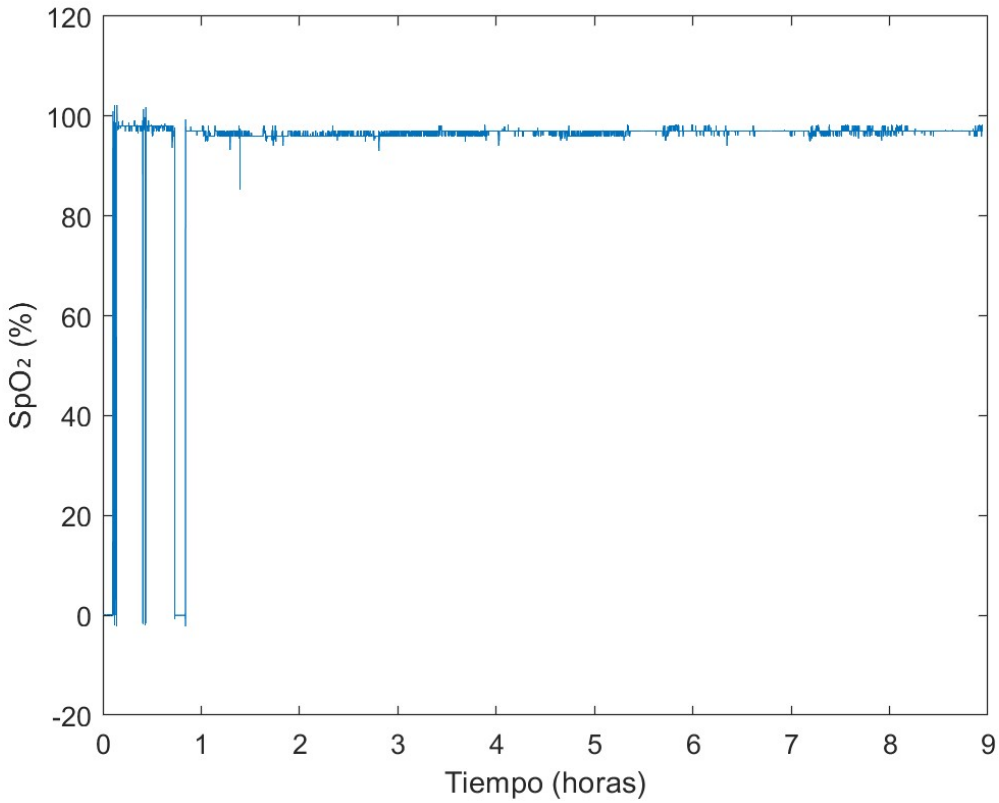

Figura 1. Señal SpO_2 de la base de datos CHAT.

1.4 ESTADO DEL ARTE DEL DIAGNÓSTICO AUTOMÁTICO DE LA AOS A PARTIR DE IMAGEN

El uso de imágenes procedentes de señales biomédicas es una técnica que se emplea actualmente en el diagnóstico de AOS. Por esta razón, existen varios estudios donde estas representaciones, que con frecuencia son espectrogramas y escalogramas, se combinan con técnicas de *machine learning* y *deep learning* (DL). La finalidad de estos procedimientos es crear herramientas nuevas para combinarlas con diferentes pruebas diagnósticas y así reducir la necesidad de recurrir solamente al empleo de la PSG convencional. En la Tabla 1 se presentan varios estudios recientes que emplean este enfoque, detallando la señal de estudio, la población, la base de datos, el método aplicado, la técnica de transformación de la señal a imagen, el método de eXplainable Artificial Intelligence (XAI) y los resultados.

Como se muestra en la Tabla 1, la mayoría de las investigaciones son en población adulta y emplean principalmente la señal de ECG para generar las imágenes de entrada a la red. Solo los estudios de Mortazavi *et al*., (2023) y Crowson *et al*., (2023) se realizan en población infantil. Por otro lado, el tipo de imagen más común son los espectrogramas obtenidos mediante la transformada de Fourier de tiempo corto (*Short-time Fourier Transform*, STFT) y los escalogramas mediante la transformada Wavelet continua (*Continuous Wavelet Transform*, CWT). Además, únicamente Mortazavi *et al*., (2023) estima el AHI para clasificar la severidad de la AOS, mientras que otros trabajos como Wang *et al.*, (2022) y Song *et al.*, (2023) clasifican primero los eventos respiratorios para después estimar el AHI. El resto de los estudios se centran principalmente en la clasificación de eventos.

Cabe destacar que un número limitado de estudios incorporan técnicas XAI. Los trabajos que las incluyen se limitan a *Gradient-weighted Class Activation Mapping* (Grad-CAM) y *Local Interpretable Model-agnostic Explanation* (LIME). Esto abre la posibilidad de la aplicación de técnicas XAI alternativas como SHAP, innovando en las explicaciones y aumentando la fiabilidad en las predicciones.

AUTOR	SEÑAL	POBLACIÓN	BASE DE DATOS	MÉTODO	TRANSFORMACIÓN DE LA SEÑAL A IMAGEN	MÉTODO XAI	RESULTADOS
(Mortazavi *et al.*, 2023)	SpO_2	Infantil	CHAT	Inception V3	Espectrograma (STFT) de segmentos de 20min (75x75)	-	4 clases: Acc=74.11%, Kappa= 0.6 ICC=0.96
(Crowson *et al.*, 2023)	NAP	Infantil	Privada	ResNet50	Escalogramas (CWT) de segmentos de 30s (512x512)	-	4 clases: Acc=70.0%
(Stewart *et al.*, 2023)	SpO_2, TH, NP, ECG	Adulta	Privada	ResNet18	Espectrograma (STFT) de segmentos de 60s	-	2 clases: Acc=94.18%, Recall=92.95%, Sp=95.4%
(Ullah *et al.*, 2023)	ECG	Adulta	Apnea-ECG	1D-CNN*, 2D-CNN, 2D-CNN + SVM	Espectrograma (STFT), Escalograma (CWT) de segmentos de 60s (299x299)	-	2 clases: Acc=98%, F1-score=97.5%

(Mashrur *et al.*, 2021)	ECG	Adulta	Apnea-ECG y UCDDB	2D-CNN	Escalograma (CWT) de segmentos de 60s (32x32)	-	2 clases: Apnea-ECG: Acc: 94.38% Se: 94.30%
(Choudhury *et al.*, 2025)	ECG	Adulta	Apnea-ECG, UCDDB y MIT-BIH	GoogLeNet, AlexNet, VGG-16, ResNet18	Escalograma (CWT) de segmentos de 30 y 60s (224x224)	LIME	2 clases: Apnea-ECG, GoogleNet: Acc: 93.85%, Se: 93.42%, Sp: 94.30%, F1-score: 93.83%
(Gupta, Bajaj and Ansari, 2022).	ECG	Adulta	Apnea-ECG	2D-CNN, ResNet50, Squeezenet	Espectrogramas de Gabor (GS) de segmentos de 60s (224x224, 227x227)	-	2 clases: 2D-CNN: Acc=94.81%, Se=94.58%, Sp=94.95%, F1-score=94.75%, AUC=98.92%
(Zhou and Kang, 2024)	ECG	Adulta	Apnea-ECG	2D-CNN	Escalograma (CWT) de segmentos de 60s (224x224)	-	2 clases: Acc=96.37%, Se=94.67%, Sp=97.44%, AUC=96%
(Linh *et al.*, 2024)	ECG	Adulta	Privada	SVM, KNN, DT, LDA	Escalograma (CWT) de segmentos de 30s	-	2 clases: Acc=98.2%, F1-score=93%
(Akter *et al.*, 2025)	ECG	Adulta	Apnea-ECG	2D-CNN	Escalograma(CWT) de segmentos de 60s (128x180)	Grad-CAM	2 clases: Acc=98.79%, Se=98.93%, Sp=98.9%
(Niroshana *et al.*, 2021)	ECG	Adulta	Apnea-ECG	2D-CNN	Escalograma (CWT), espectrograma (STFT) e imágenes fusionadas de segemntos de 60s (128x128)	-	2 clases: Acc=92.4%, Recall=92.3%, Sp=92.6%
(Lin *et al.*, 2024)	ECG	Adulta	Apnea-ECG	EfficientNet + Mecanismo de atención	Escalograma (CWT) de segmentos de 1,2 y 3 min (224x224)	-	2 clases: Acc=93.44%, Se= 88.9%, Sp=96.2%
(Nasifoglu and Erogul, 2021)	ECG	Adulta	HomePAP, ABC	2D-CNN, ResNet18, GoogleNet, AlexNet	Escalograma (CWT) y espectrograma (STFT) de segmentos de 30s (224x224, 227x227)	Grad-CAM	2 clases: CWT, 2D-CNN: Acc=82.30%, Se=83.22%, Sp=82.27% STFT, 2D-CNN: Acc=80.13%, Se=81.99%, Sp=77.25%
(Yu *et al.*, 2021)	ECG	Adulta	Apnea-ECG	2D-CNN + Bi-LSTM + Mecanismo de atención	Espectrogramas (STFT) de segmentos de 60s (20x92,26x9233x92)	-	2 clases: Acc=87.09%, Se=77.96%, Sp=91.74%, F1 score=81.61%
(Wang, Koprinska and Jeffries, 2023)	FA, señal torácica y señal abdominal	Adulta	MESA	1D- CNN, ConvLSTM, 1D-CNN-LSTM, 2D-CNN-LSTM	Escalograma (CWT) de segmentos de 30 y 90s (128x128)	-	2 clases: 2D-CNN-LSTM: Acc=83.90%, Se=81.21%, Sp=86.59%
(Wu *et al.*, 2021)	FA	Adulta	Privada	2D-CNN	Espectrograma (STFT) de segmentos de 15s	-	3 clases: Acc= 91.23%, Se= 90.81%, Sp=90.59%

(Moradhasel *et al.*, 2023)	EMG Mentón	Adulta	Privada	EfficientNet, MobileNet, VGG-16, MLP	Espectrogramas (STFT) de segmentos de 1s (224x224)	-	2 clases: EfficientNet: Acc=99.1%, MobileNet: Acc= 94.5% VGG-16: Acc= 94.8% MLP: Acc= 76.8%
(Tanci and Hekim, 2025)	EEG	Adulta	Apnea-ECG	ResNet64, YOLOv5, YOLOv8	Espectrograma (STFT) de segmentos 30s (224x224)	-	4 clases: YOLOv8: Acc= 93.7% ResNet64: Acc=93%. YOLOv5: Acc=88.2%
(Romero *et al.*, 2022)	Sonido de respiración	Adulta	Privada	2D-CNN	Espectrograma de Mel	-	2 clases: Se=79%, Sp=80%
(Wang *et al.*, 2022)	Sonido de respiración	Adulta	Privada	2D-CNN	Espectrograma de Mel	-	2 clases: Acc=81%, Se=78%
(Song *et al.*, 2023)	Sonidos de respiración	Adulta	Privada	CNN+ ResNet18+ XGBoost	Espectrograma de Mel	-	2 clases: Acc: 83.44%, Recall=8527%

*Tabla 1. Estado del arte del empleo de imagen junto técnicas de DL y ML para el diagnóstico de la AOS *Redes neuronales convolucionales (Convolutional Neural Network, CNN).*

1.4.1 DEEP LEARNING

El *deep learning* (DL) surge como una rama del *machine learning.* Tradicionalmente se ha basado en redes neuronales profundas capaces de aprender patrones complejos de los datos en crudo (*raw data*) sin necesidad de una previa extracción manual de información o características, enfoque típico en metodologías previas (Lecun, Bengio and Hinton, 2015; Janiesch, Zschech and Heinrich, 2021).

El DL no solo tiene la ventaja de aprender de forma automatizada las características, incluso aunque éstas no sean conocidas por el ser humano, sino que también presenta una gran flexibilidad adaptándose a tareas complejas y siendo eficaz con grandes volúmenes de datos (Janiesch et al., 2021). Esto permite que el aprendizaje alcance altos niveles de abstracción (Janiesch et al., 2021; Lecun et al., 2015).

Se aplica DL para una variedad de tareas como el procesado de imagen, el procesado de lenguaje natural o el reconocimiento de voz y audio entre otras, incluyendo un importante rol en el área de la ciencia y, específicamente, la medicina (Lecun, Bengio and Hinton, 2015; Janiesch, Zschech and Heinrich, 2021).

El DL se presenta como una técnica muy relevante en el ámbito de la imagen biomédica, debido a su capacidad para emular el procesado visual del cerebro humano. Este tipo de técnicas es capaz de identificar y aprender progresivamente desde estructuras básicas, como bordes y esquinas, hasta patrones complejos presentes en las imágenes (Bhatt *et al.*, 2021a; Janiesch, Zschech and Heinrich, 2021). En el ámbito clínico, el DL se aplica sobre imágenes generadas por dispositivos médicos, como la resonancia magnética y la tomografía computarizada, con el fin de extraer automáticamente características relevantes. Esto permite el diagnóstico automatizado de enfermedades, la segmentación de estructuras patológicas y la ayuda a la decisión médica (Bhatt *et al.*, 2021a).

Las ventajas del DL en el análisis de imágenes también abren la posibilidad de aplicar estas técnicas a imágenes generadas a partir de señales fisiológicas obtenidas mediante otros dispositivos, como ECG, EEG, SpO_2 o EMG. Esta línea de investigación está centrada en la transformación de las señales al dominio tiempo-frecuencia (mapas frecuenciales) y la aplicación de las arquitecturas de DL para imagen. Existe una amplia variedad de arquitecturas pre-entrenadas de redes neuronales destinadas al análisis de imágenes. Las redes más empleadas en estas tareas son diferentes tipos de la arquitectura ResNet (ResNet50, ResNet18, ResNet64), EfficientNet, Inception y varias redes 2D-CNN modificadas (Mashrur *et al.*, 2021; Nasifoglu and Erogul, 2021; Niroshana *et al.*, 2021; Wu *et al.*, 2021; Yu *et al.*, 2021; Gupta, Bajaj and Ansari, 2022; Crowson *et al.*, 2023; Moradhasel *et al.*, 2023; Mortazavi *et al.*, 2023; Stewart *et al.*, 2023; Ullah *et al.*, 2023; Wang, Koprinska and Jeffries, 2023; Lin *et al.*, 2024; Zhou and Kang, 2024; Akter *et al.*, 2025; Choudhury *et al.*, 2025; Tanci and Hekim, 2025). El objetivo que se quiere lograr con el empleo de estas técnicas es contribuir al diagnóstico de la AOS, ya sea dando una estimación del AHI, prediciendo la presencia o ausencia de la enfermedad o clasificando la AOS en diferentes severidades.

1.4.2 EXPLAINABLE ARTIFICIAL INTELLIGENCE

Una de las principales desventajas asociadas al DL es el denominado fenómeno de la *caja negra*, el cual hace referencia a la dificultad en la interpretación de la toma de decisiones de la red (Adadi and Berrada, 2018). Esto quiere decir que no se puede saber la razón detrás de las predicciones de los modelos, lo que disminuye considerablemente su comprensión y, por tanto, la potencial confianza depositada en ellos (Lecun et al., 2015).

La inteligencia artificial explicable (*eXplainable Artificial Intelligence*, XAI) ha emergido en los últimos años como solución a esta falta de transparencia en cuanto a las decisiones tomadas por los modelos de aprendizaje automático. Sin embargo, esto supone un desafío, ya que los modelos con rendimientos más elevados suelen ser los más opacos y, al contrario, los más interpretables son los de menor rendimiento potencial, por lo que hay que encontrar un cierto equilibrio (Gunning *et al.*, 2019; Samek *et al.*, 2019). No existe un único método de XAI, sino que es un conjunto de técnicas que ayudan a interpretar qué partes de los datos son importantes en las decisiones, predicciones o estimaciones realizadas por los modelos (Adadi and Berrada, 2018). En el ámbito del análisis de imágenes, la aplicación de XAI resulta especialmente valiosa, ya que algunas de estas técnicas permiten generar mapas de calor que destacan las regiones de píxeles con mayor influencia en la salida del modelo (Gunning *et al.*, 2019; Samek *et al.*, 2019).

Algunos trabajos recientes han comenzado a explorar estas posibilidades para la detección de AOS. Por ejemplo, en el caso de Akter *et al*. y Huseyin Nasifoglu *et al*. se emplea Grad-CAM con la señal ECG en adultos, una herramienta que calcula los gradientes de la última capa de activación de la red para crear mapas de calor que indican regiones de la imagen más o menos influyentes en la predicción de la CNN (Nasifoglu and Erogul, 2021; Akter *et al.*, 2025). Otra técnica de XAI, es LIME, empleada en el

estudio de M. Choudhury *et al*. con la señal de ECG en adultos. Esta técnica genera perturbaciones de las imágenes de entrada para observar las variaciones en la salida y generar así también mapas de calor con las explicaciones sobre las imágenes (Choudhury *et al.*, 2025).

Pese a su relevancia, son pocos los estudios en estas áreas que implementan técnicas de XAI para completar la interpretación de los resultados obtenidos, por lo que su empleo en este trabajo se presenta como algo novedoso.

1.5 COMPARACIÓN Y ELECCIÓN DEL MÉTODO A IMPLEMENTAR

Para la transformación de señales fisiológicas en representaciones visuales, se ha optado por emplear la STFT, con el objetivo de convertir las señales de SpO_2 en mapas tiempo-frecuencia (espectrogramas) a partir de los cuales la CNN pueda extraer características. Si bien tanto la STFT como la CWT han demostrado ser técnicas eficaces y ampliamente utilizadas en este tipo de tareas, la STFT fue seleccionada por su menor complejidad computacional y su capacidad de generar resultados comparables a los de la CWT

Respecto al modelo de DL, se ha optado por un enfoque de *transfer learning* mediante la arquitectura ResNet, la cual ha sido empleada en diferentes trabajos previos con imágenes alcanzando altos rendimientos (Crowson et al., 2023; Stewart et al., 2023; Choudhury et al., 2025; Gupta, Bajaj and Ansari, 2022; Nasifoglu and Erogul, 2021; Song et al., 2023). En particular, ResNet50 se ha empleado con imagen obteniendo buenos resultados con imagen, tanto derivada de la señal de SpO_2 como otros tipos de imagen médica, como las retinografías (Gupta, Bajaj and Ansari, 2022; Crowson *et al.*, 2023; Herrero-Tudela *et al.*, 2025).Haga clic o pulse aquí para escribir texto. Por ejemplo, Crowson et al., (2023) utiliza hasta 9 redes diferentes para la apnea infantil (varios tipos de DenseNet, ResNet y VGG) alcanzando los mejores resultados con ResNet50. Con el fin de evaluar el impacto del pre-entrenamiento sobre distintas distribuciones de color, se utilizarán dos versiones de ResNet50: una pre-entrenada con el conjunto de datos *ImageNet* (imágenes generales) y otra pre-entrenada con imágenes de retina, cuya similitud cromática con los espectrogramas generados podría influir en el rendimiento del modelo.

Finalmente, dado que las técnicas de XAI aún no han sido ampliamente exploradas en esta área, su incorporación en el presente trabajo resulta de especial interés. Se ha seleccionado el método *Shapley Aditive eXplanations* (SHAP) debido a que es diferente a las técnicas abordadas en trabajos previos (LIME y Grad-CAM). Esto se hace con el objetivo de encontrar nuevas explicaciones de las imágenes, tratando de dar un enfoque más novedoso a la transparencia del modelo.

1.6 HIPÓTESIS Y OBJETIVOS

Este Trabajo Fin de Grado (TFG) presenta como hipótesis que el DL, junto con técnicas XAI, son capaces de dar un diagnóstico automático y clínicamente relevante de la AOS en niños a partir de espectrogramas procedentes de la señal de oximetría. El objetivo general del trabajo es desarrollar un modelo de DL explicable que estime el AHI y dar una estadificación de la enfermedad a partir de él. Esta estimación se llevará a cabo mediante una tarea de regresión aplicada sobre espectrogramas generados a partir de registros de SpO_2 extraídos de la base de datos semipública *Childhood Adenotonsillectomy Trial* (CHAT).

Partiendo del objetivo general, los objetivos específicos para desarrollar este trabajo son:

1. Llevar a cabo una revisión del estado del arte sobre el empleo de imagen procedente de señales biomédicas junto con técnicas de DL cuya población objetivo sea pacientes con apnea.
2. Convertir las señales de SpO_2 de la base de datos CHAT en espectrogramas mediante el método elegido.
3. Realizar *transfer learning* con la red neuronal seleccionada para la estimación del AHI y la posterior estadificación de la enfermedad.
4. Realizar *transfer learning* con la red neuronal seleccionada pre-entrenada con imágenes de retina para la estimación del índice AHI y la posterior estadificación de la enfermedad a partir de este.
5. Introducir ajustes en la arquitectura de las redes neuronales empleadas, con el fin de optimizar sus resultados.
6. Aplicar técnicas XAI para interpretar los resultados obtenidos.
7. Comparar y extraer conclusiones de los resultados obtenidos en el estudio.

1.7 PLAN DE TRABAJO Y ESTRUCTURA DEL TFG

1.7.1 PLAN DE TRABAJO

Para alcanzar todos los objetivos, en este trabajo se ha llevado a cabo el siguiente plan de trabajo dividido en diferentes fases:

- **Fase 1:**
    - Comprensión de los procedimientos y reglas del diagnóstico de apnea en niños a partir del manual de la AASM de 2007 y sus actualizaciones realizadas en 2018.
    - Lectura de documentos específicos de la apnea del sueño para llevar a cabo un aprendizaje sobre las características de la enfermedad en población infantil.
    - Realización de cursos DL básicos y cursos de DL centrados en las CNN acompañados de ejercicios prácticos.
    - Realización de cursos de Python para perfeccionar su aprendizaje.

    - Realización de una revisión bibliográfica de los estudios publicados que combinen técnicas de *machine learning* o DL para el diagnóstico de apnea a partir de imagen procedente de señal biomédica.
- **Fase 2:**
    - Elección y aplicación del método de transformación de las señales SpO_2 de la base CHAT en imagen.
    - Elección y aplicación de la red pre-entrenada general sobre las imágenes de SpO_2 para la estimación del AHI y la estadificación de la enfermedad a partir de este.
    - Elección y aplicación de la red pre-entrenada con imágenes de retina sobre las imágenes de SpO_2 para la estimación del AHI y la estadificación de la enfermedad a partir de este.
    - Cálculo de diferentes métricas de rendimiento para medir y comparar la exactitud de los resultados.
    - Implementación de cambios en la arquitectura de la red para lograr mejorar resultados.
    - Elección del mejor modelo entre todos los obtenidos.
- **Fase 3:**
    - Elección y aplicación del método XAI con el mejor modelo obtenido para lograr la interpretación de las decisiones sobre los datos.
    - Cálculo del coeficiente de correlación intraclase (*Intraclass Correlation Coefficient*, ICC) del AHI real y predicho por el mejor modelo obtenido.
    - Observación de la concordancia entre las predicciones AHI del mejor modelo obtenido y los AHI reales partir de un diagrama Bland-Altman.
- **Fase 4:**
    - Análisis y extracción de conclusiones a partir de todos los resultados presentados.

1.7.2 ESTRUCTURA DEL TFG

El resto del TFG se ha organizado de la siguiente forma:

El *Capítulo 2: Sujetos e imágenes de estudio,* presenta las características de la base de datos CHAT, de la cual se extraen las señales de SpO_2 para este estudio. También se explica en qué consiste la separación de los datos en tres grupos diferentes, así como su finalidad dentro del diseño experimental.

El *Capítulo 3: Metodología,* detalla todos los métodos empleados para llevar a cabo este trabajo. Se describe la transformación de las señales en espectrogramas mediante STFT, se introducen los fundamentos de las CNN, y se explica la arquitectura ResNet y sus variantes. También se justifica el uso de la técnica SHAP como herramienta de explicabilidad. Una vez introducidas las técnicas de base, se describen las arquitecturas empleadas. Por una parte, la red ResNet50 pre-entrenada de *ImageNet* y, por otra parte, ResNet50 pre-entrenada con imágenes de retina. Finalmente, se describen las métricas

utilizadas para evaluar el rendimiento del modelo tanto en tareas de regresión como de clasificación multiclase.

El *Capítulo 4: Resultados* presenta, la optimización de hiperparámetros en ambos modelos (ResNet50 y ResNet50-Retina). Se incluyen métricas cuantitativas y representaciones gráficas. Se presentan también los análisis interpretativos con SHAP, ilustrando visualmente las decisiones del modelo para las mejores y peores predicciones.

El *Capítulo 5: Discusión de los resultados,* discute los resultados obtenidos. Primero se aborda la clasificación de la apnea en 4 clases, explicando lo que se ha obtenido para cada uno de los mejores modelos y realizando una comparación entre estos. Asimismo, se comenta y compara el rendimiento en la estimación del AHI.

El *Capítulo 6: Conclusiones y líneas futuras,* se centra en las principales contribuciones a la investigación y las conclusiones más relevantes extraídas de este trabajo. Tras esto, se establecen las posibles líneas futuras poniendo fin al trabajo.

CAPÍTULO 2: SUJETOS E IMÁGENES DEL ESTUDIO

2.1 BASE DE DATOS

La base de datos utilizada en este trabajo corresponde al *Childhood Adenotonsillectomy Trial* (CHAT), un ensayo clínico aleatorizado desarrollado entre los años 2007 y 2011 por un consorcio de investigadores de diversas instituciones. Esta base de datos se encuentra disponible públicamente a través del enlace https://sleepdata.org/datasets/chat.

El objetivo principal del estudio clínico fue evaluar si la realización temprana de una adenoamigdalectomía (AT) en niños con AOS se asociaba con una mejora significativa en la calidad de vida, así como en las funciones cognitivas, conductuales y metabólicas (Marcus *et al*., 2012).

La base de datos está compuesta por un total de 1.638 registros correspondientes a niños estadounidenses con edades comprendidas entre los 5 y los 9,9 años. Los participantes fueron asignados de manera aleatoria a uno de dos grupos experimentales: el grupo de *early Adenotonsillectomy* (eAT), en el que los niños fueron sometidos a cirugía temprana, y el grupo *Watchful Waiting with Supportive Care* (WWSC), que no recibió intervención quirúrgica inmediata, sino seguimiento clínico y cuidados relacionados con la salud del sueño (*Childhood Adenotonsillectomy Trial*, no date).

La base de datos CHAT recoge información del sueño de estos niños mediante una prueba de PSG. La información de la PSG incluye la señal de SpO_2, el AHI y etiquetas con la localización temporal precisa de los eventos de apnea e hipopnea siguiendo las directrices de la AASM. La señal de saturación también se empleó para calcular el ODI. Asimismo, se eliminaron los casos de niños con hipoxemia severa con SpO_2 <90% durante >2% del tiempo (Marcus *et al.*, 2013).

2.2 PARTICIÓN DEL CONJUNTO DE DATOS

Debido a que en este trabajo se han empleado redes neuronales, es imprescindible realizar una división de los datos en los siguientes tres grupos: entrenamiento, validación y test. El objetivo es evitar el sobreajuste, es decir, que la red aprenda de forma específica los patrones de los datos de entrenamiento y sea incapaz de generalizar cuando se empleen datos que no ha observado previamente (Janiesch, Zschech and Heinrich, 2021)

Por un lado, se necesita un conjunto de entrenamiento con la mayor cantidad de datos posible, ya que es con los que el modelo va a ajustar sus parámetros (Bengio, Goodfellow and Courville, 2015). Estos datos son pasados a la red junto a sus etiquetas para que pueda aprender relaciones entre ellos y encontrar patrones característicos de las imágenes. El

ajuste se realiza mediante la minimización de una función de pérdida, la cual cuantifica el error y puede adoptar distintas formas según la naturaleza del problema, como el error cuadrático medio (*mean squared error*), la entropía cruzada (*cross-entropy*) o la entropía cruzada categórica (*categorical cross-entropy*), entre otras (Bengio, Goodfellow and Courville, 2015).

El conjunto de validación tiene como finalidad controlar la evolución del entrenamiento, mediante el ajuste de hiperparámetros. Aunque este conjunto no interviene directamente en el ajuste de los pesos de la red, supervisa el modelo influyendo en su configuración y estimando su capacidad de generalización (Janiesch, Zschech and Heinrich, 2021).

Por último, el conjunto de test se utiliza para comprobar el rendimiento del modelo desarrollado. Se destina una parte de las imágenes para poder evaluar la red en imágenes distintas a las que se han usado durante el entrenamiento y la validación, obteniendo así métricas de rendimiento no sesgadas (Bengio, Goodfellow and Courville, 2015; Janiesch, Zschech and Heinrich, 2021).

En la Tabla 2 se presentan los datos sociodemográficos y clínicos correspondientes a los tres subconjuntos generados para este estudio: entrenamiento, validación y test. La partición de los datos se realizó de forma aleatoria a nivel de sujeto, es decir, cada participante fue asignado exclusivamente a uno de los subconjuntos para evitar cualquier tipo de solapamiento entre ellos (Jiménez-García *et al.*, 2024). A partir de un total de 1.639 imágenes, se asignaron 1.006 al conjunto de entrenamiento, 326 al conjunto de validación y 306 al conjunto de prueba, lo que corresponde a proporciones del 61,42 %, 19,90 % y 18,68 %, respectivamente. Esta distribución se considera adecuada, ya que permite destinar la mayoría de los datos al entrenamiento del modelo, mientras que los subconjuntos de validación y test conservan un tamaño suficiente para evaluar de manera fiable su rendimiento y capacidad de generalización. Asimismo, se comprobó que no existían diferencias estadísticamente significativas ($p > 0{,}01$) entre los tres grupos en variables clínicas relevantes como la edad, el sexo, el índice de masa corporal ajustado por edad (*BMI z-score*) y el AHI (Jiménez-García *et al.*, 2024).

Finalmente, en la Figura 2 se muestra un esquema del proceso llevado a cabo: obtención de las señales de SpO_2 de la base de datos CHAT, su transformación en espectrogramas y la posterior división de los datos en tres grupos.

	Entrenamiento	Validación	Test
Sujetos (n)	1006 (61.42%)	326 (19.90%)	306 (18.68%)
Edad (years)	7 [6; 8]	7 [6; 8]	6.9 [6; 8]
Mujeres (n)	520 (51.7%)	168 (51.5%)	168 (54.9%)
Varones (n)	471 (46.8%)	156 (47.9%)	134 (43.8%)
BMI (kg/m²)	17.4 [15.6; 21.7]	17.1 [15.4; 21.8]	17.6 [15.7; 21.7]
AHI (eventos/h)	2.6 [1.1; 5.9]	2.4 [1.2; 5.8]	2.3 [1.1; 6.2]
No AOS (n)	219 (21.8%)	69 (21.2%)	67 (21.9%)
AOS leve (n)	496 (49.3%)	168 (51.5%)	148 (48.4%)
AOS moderada (n)	160 (15.9%)	44 (13.5%)	49 (16.0%)
AOS severa (n)	131 (13.0%)	45 (13.8%)	42 (13.7%)

Tabla 2. Tabla con datos sociodemográficos de la base de datos CHAT.

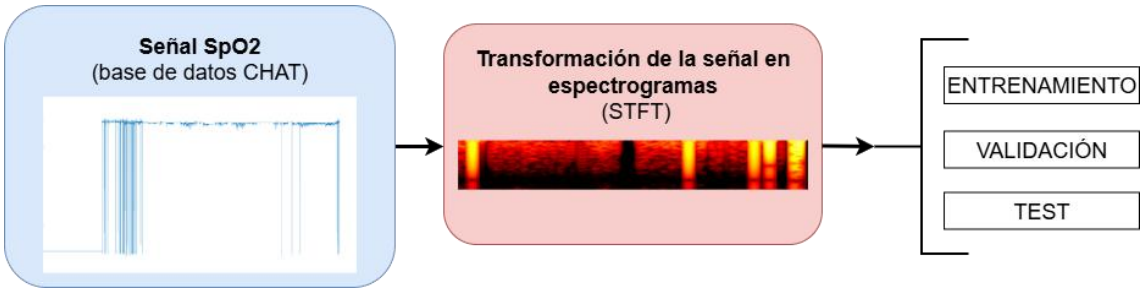

Figura 2. Esquema del proceso de transformación de la señal a imagen y separación de los datos.

CAPÍTULO 3: METODOLOGÍA

3.1 TRANSFORMACIÓN DE LA SEÑAL EN IMAGEN

En el presente estudio se han utilizado imágenes generadas a partir de señales de SpO_2 como entrada a la red neuronal. Para la obtención de estas representaciones visuales, se optó por la aplicación de la STFT por tratarse de una técnica robusta y ampliamente validada en el análisis de señales biomédicas, particularmente adecuadas para señales no estacionarias como las fisiológicas. Su cálculo es el siguiente (Ecuación 2) (Stankovic and Dakovic, 2013; Wacker and Witte, 2013):

$$\mathrm{STFT}(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]\; w[m-n]\, e^{-j\omega m} \tag{2}$$

Donde $x[m]$ es la señal en tiempo discreto, $w[m-n]$ es la ventana centrada en el tiempo m, y $e^{-j\omega m}$ es la exponencial compleja que transforma al dominio frecuencial. La ventana se desplaza a lo largo del tiempo y permite dividir la señal en segmentos temporales para analizar cómo evoluciona en frecuencia. Existe un compromiso entre la resolución en tiempo y la resolución en frecuencia, de manera que una ventana de mayor duración resulta en una mejor resolución en frecuencia y peor en tiempo, mientras que con una ventana más corta ocurre lo contrario (Stankovic and Dakovic, 2013).

La conversión de las imágenes de SpO_2 en espectrogramas se llevó a cabo mediante el uso de MATLAB©. Se creó un script para procesar secuencialmente las señales de la base CHAT que se encontraban almacenadas en formato *.mat*.

Se empleó una función para la generación de los espectrogramas que aplica la STFT para obtener la representación de la señal en el dominio tiempo-frecuencia. Los espectrogramas se calcularon con la misma frecuencia de muestreo, ya que todas las señales se presentan remuestreadas a 1 Hz, con un número constante de puntos de la STFT, el cual fue 1024. En este contexto, el número de puntos de la STFT se refiere a las componentes frecuenciales calculadas por cada ventana temporal (Wacker & Witte, 2013). Por esta razón, el empleo de 1024 puntos implica una buena resolución espectral (Okumura, 2011).

El solapamiento indica la cantidad de muestras comunes entre dos ventanas consecutivas. En el análisis, la ventana se desplazó una muestra por vez, lo que implica un solapamiento elevado. Cada ventana tiene las mismas muestras que la anterior salvo una, esta redundancia de información permite observar mejor la evolución temporal de las componentes frecuenciales (Okumura, 2011).

Se definió el mapa de color '*Hot'* en la representación debido a la similitud de tonalidades con las imágenes de retina. Como resultado se obtuvo una matriz de datos correspondiente a potencia (PSD) de la señal. Se convirtieron sus unidades a decibelios (dB) y se

estableció un límite superior e inferior de [-30,50] dB para homogeneizar los valores del mapa y mejorar su visualización.

Las imágenes fueron generadas en el mismo tamaño que la matriz de datos del espectrograma, por lo que la dimensión resultante de cada una de ellas fue diferente, coincidiendo todas en alto (eje de las frecuencias) 513 píxeles y teniendo un largo (eje temporal) variable entre 30.000 y 50.000 píxeles. Por lo tanto, como resultado del código implementado, se obtuvieron 1.638 espectrogramas en formato .png, cada uno correspondiente a una las señales completas.

Tras pruebas comparativas utilizando ventanas de 1 y 2 minutos, se concluyó que la ventana de 2 minutos ofrecía mejores resultados. Esto se debe a que, al trabajar en un rango de frecuencias bajas de 0-0.5 Hz con variaciones lentas, una ventana más larga mejora la resolución espectral para la detección precisa de eventos sin comprometer la información temporal relevante. En la Figura 3 se muestra una señal de SpO2 de la base de datos CHAT con el mapa de color '*Hot'* empleado. Este varía desde el color negro para potencias bajas hasta el amarillo para las más altas.

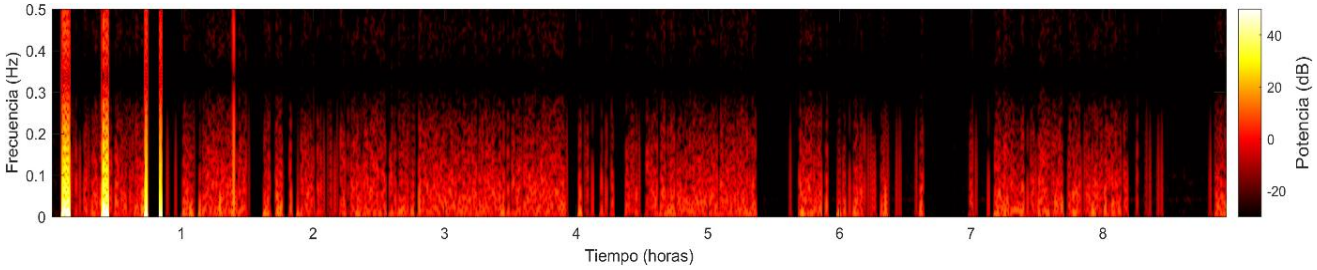

Figura 3. Espectrograma de señal SpO_2 con mapa de color Hot y ventana de 2 minutos.

3.2 REDES NEURONALES CONVOLUCIONALES

Las CNNs constituyen una clase de arquitecturas de aprendizaje profundo especialmente eficaces en tareas que implican la interpretación de imágenes, aplicándose con éxito en problemas de regresión, clasificación y segmentación (Wu, 2017; Bhatt *et al*., 2021b). En el caso específico de las imágenes digitales, los datos de entrada se representan como matrices numéricas, cuyos valores oscilan entre 0 y 255, y corresponden a la intensidad de los píxeles de uno o varios canales (Wu, 2017; Bhatt *et al.*, 2021b).

El núcleo funcional de estas redes lo constituyen las capas convolucionales, que aplican filtros (*kernels*), es decir, matrices de pequeña dimensión que se desplazan por la imagen realizando multiplicaciones y sumas para generar mapas de características, que codifican patrones espaciales relevantes para el modelo (Figura 4) (Wu, 2017; Bhatt *et al.*, 2021b).

Tanto la información como el tamaño de los mapas generados dependen de diferentes parámetros como los pesos y los sesgos que emplea en las operaciones matemáticas. Por otro lado, el número de filtros por capa define su profundidad y permite capturar diferentes características. Asimismo, son importantes el *stride*, o paso, que determina la

cantidad de desplazamiento del filtro sobre la imagen; y el *padding*, que consiste en añadir ceros alrededor de los bordes de la imagen con el fin de conservar sus dimensiones originales tras la convolución (Wu, 2017; Bhatt *et al.*, 2021b).

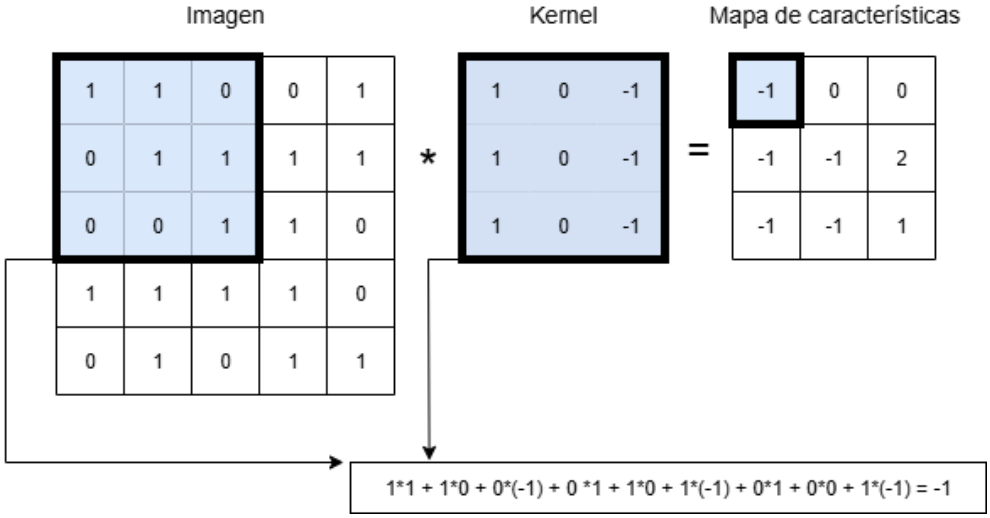

Figura 4. Ejemplo de una operación de convolución con kernel de 3x3 sobre una imagen de 5x5 píxeles.

A la salida de cada capa convolucional se aplican funciones de activación que aumentan la capacidad de aprendizaje de las redes. Existen diferentes tipos de funciones de activación como la sigmoide, la tangente hiperbólica o la *Rectified Linear Unit* (ReLU). Particularmente, esta última, ReLU, es adecuada debido a que añade no linealidades que son útiles para detectar los patrones complejos que presentan las imágenes. A diferencia de otras funciones, evita el desvanecimiento del gradiente y es más fácil de calcular. ReLU convierte a 0 los números negativos y deja inalterados los positivos, respondiendo a la siguiente ecuación (Ecuación 3) (Wu, 2017; Bhatt *et al.*, 2021b).

$$ReLU(x) = max\ (0, x) \tag{3}$$

Las capas de *pooling* permiten reducir la dimensionalidad espacial de los mapas de características. Esto permite reducir el número de operaciones a realizar y minimizar el riesgo de sobreajuste, (Wu, 2017; Bhatt *et al.*, 2021b). Existen diferentes formas de realizar *pooling* como por ejemplo el *max pooling* que toma el píxel con valor más alto de todos los pertenecientes a una pequeña región, conservando así las activaciones fuertes. Otro ejemplo es el *average pooling,* que, en vez del máximo, toma el promedio de todos, conservando una información más general (Wu, 2017; Bhatt *et al.*, 2021b).

Además, es posible la integración de capas de *batch normalization,* que normalizan las salidas de capas previas mejorando la estabilidad y la velocidad del entrenamiento (Wu, 2017). Por otra parte, están las capas *fully conected*, que toman los datos de la última capa de *pooling* y los conecta con cada una de las neuronas de salida. Esto va acompañado de una función de activación que genera la salida y se adecua a la tarea a realizar (Wu, 2017).

El proceso de aprendizaje en una CNN se lleva a cabo mediante un ciclo iterativo que comprende dos fases principales: la propagación hacia adelante (*forward pass*), en la que

los datos atraviesan la red para generar una predicción, y la retropropagación del error (*backpropagation*), donde se calcula la función de pérdida comparando la predicción con la etiqueta real y se actualizan los pesos de los filtros con el objetivo de minimizar dicho error (Wu, 2017).

3.2.1 RESNET

La *Residual Network* (ResNet) es una arquitectura de red neuronal profunda diseñada inicialmente para trabajar con imágenes. Ha demostrado una alta eficacia en diversas aplicaciones, particularmente en la medicina, pudiendo dar diagnósticos tempranos, segmentando estructuras patológicas o clasificando enfermedades (Xu, Fu and Zhu, 2023).

Esta arquitectura se caracteriza por estar compuesta de bloques residuales cuya finalidad es solucionar el desvanecimiento del gradiente. Este es un problema que presentan las redes neuronales muy profundas a medida que se añaden nuevas capas y que resulta en ocasiones en un rendimiento menor del deseado (He *et al.*, 2016; Xu, Fu and Zhu, 2023).

Cada bloque residual está compuesto por dos capas convolucionales de tamaño 3×3 y dos capas de *batch normalization* con función de activación ReLU. En esta unidad residual hay una conexión entre la entrada y la salida llamada *skip connection,* que permite sumarlas si su dimensión es la misma (Figura 5). En el caso en que la salida tenga un tamaño diferente, se realiza una operación adicional para igualarlas, que puede consistir en un mapeo lineal con una convolución de 1×1 o un *padding* con ceros. La unión de la entrada y la salida del bloque como se indica en la Ecuación 4, siendo x la entrada, $F(x, W)$ la salida, estabiliza el entrenamiento facilitando el flujo del gradiente (Xu, Fu and Zhu, 2023).

$$y = f(x + F(x, W)) \tag{4}$$

La arquitectura ResNet es adecuada para enfoques de *transfer learning* en el ámbito del procesamiento de imágenes médicas. Esto se debe a la disponibilidad de modelos ResNet pre-entrenados sobre conjuntos de datos de gran escala, como *ImageNet*. Esto permite aprovechar características de bajo nivel, como bordes o texturas, comunes a la mayoría de las imágenes y aprendidas en más de un millón de observaciones. Los pesos de estos modelos pre-entrenados son comúnmente utilizados como punto de partida para otras tareas diferentes en las que se tiene menor número de datos etiquetados (He *et al.*, 2016; Xu, Fu and Zhu, 2023).

ResNet presenta diversas variantes, diferenciadas por la profundidad de la red, como ResNet-18, ResNet-50 o ResNet-152. La elección de una u otra variante depende de la complejidad del problema y de los recursos computacionales disponibles. Las redes menos profundas, como ResNet-18, resultan más apropiadas para conjuntos de datos pequeños o problemas con menor complejidad, mientras que las variantes más profundas son más eficaces en contextos que requieren una mayor capacidad de abstracción. A partir

de ResNet-50, la arquitectura incorpora los denominados bloques *bottleneck*, que consisten en variantes de los bloques residuales con tres capas convolucionales (1×1, 3×3 y 1×1). El objetivo de esto es optimizar las pruebas reduciendo el número de parámetros (He *et al.*, 2016; Xu, Fu and Zhu, 2023).

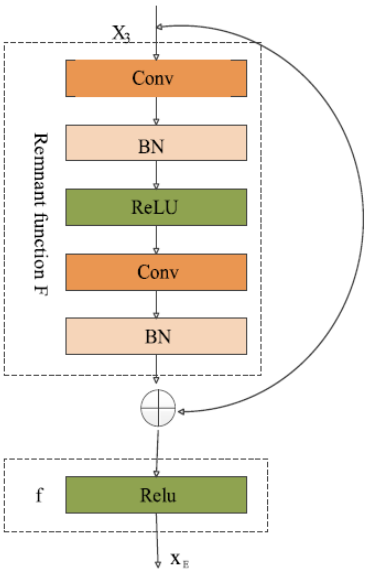

Figura 5. Esquema de un bloque residual de Resnet (Xu, Fu and Zhu, 2023).

3.3 SHAPLEY ADDITIVE EXPLANATIONS

Las técnicas de XAI han surgido con el propósito de acabar con el fenómeno de *caja negra* inherente al DL. Estas técnicas buscan proporcionar interpretaciones comprensibles sobre el funcionamiento interno de los modelos y sus predicciones (Salih *et al.*, 2024).

Una de las metodologías más consolidadas dentro del ámbito XAI es *SHapley Additive Explanations* (SHAP), un enfoque que se fundamenta en la teoría de juegos cooperativos de Shapley, que busca repartir una "recompensa" entre los jugadores en función de su contribución. En este contexto los jugadores son las características y la recompensa es la predicción, de modo que se trata de averiguar cuánto contribuye cada característica en la estimación dada por el modelo (Salih *et al.*, 2025).

SHAP es un método local, ya que permite explicar las decisiones del modelo sobre observaciones específicas; es agnóstico, dado que puede aplicarse independientemente de la arquitectura empleada; y *post-hoc*, pues se utiliza una vez que el modelo ha sido entrenado (Salih *et al.*, 2025).

Para poder medir la importancia de cada característica (en este caso, cada píxel) se emplea el valor SHAP que cumple diferentes propiedades. Debe darse una 'exactitud local', es decir, que la suma de todos los valores SHAP sea igual al resultado del modelo menos una referencia base. También debe cumplirse la 'ausencia', que implica que, si una característica no está presente, su contribución es cero. Por último, debe tener

'consistencia', que exige que el valor de una característica no disminuya si ha contribuido en la predicción (Lundberg, Allen and Lee, 2017).

En la técnica SHAP, en cada evaluación del modelo se da una combinación de características. Para ello se realiza un enmascaramiento que suprime una o más características en la entrada para comprobar cuál es su contribución. Las características ausentes no toman valor 0, sino que se sustituyen por valores de referencia resultantes de un mapeo aleatorio del espacio de entrada (Lundberg, Allen and Lee, 2017).

El valor SHAP refleja el efecto de cada característica de las imágenes al ser incluida en diferentes conjuntos de características para realizar la predicción del modelo. Esto se puede representar de forma general con la siguiente fórmula (Lundberg, Allen and Lee, 2017):

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\ (|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \tag{5}$$

donde se calcula la contribución de la característica i al modelo comparando escenarios en los que está presentes y en los que no. S es un subconjunto del resto de características, F el conjunto total y f es la función que aplica el modelo a los datos de entrada (Lundberg, Allen and Lee, 2017).

Además, para llevar a cabo esta técnica es imprescindible el uso de un valor base. Este valor se calcula como la predicción del modelo cuando no se conoce información de ninguna característica ($S = \emptyset$) y sirve como punto de partida para calcular los valores SHAP (Lundberg, Allen and Lee, 2017).

3.4 IMPLEMENTACIÓN DE LOS MÉTODOS

En este trabajo se ha implementado un código de *Python* con el objetivo de implementar una tarea de regresión orientada a la estimación del AHI. Esta estimación constituye la base para el diagnóstico automatizado y la clasificación en grados de severidad de la AOS pediátrica, a partir de imágenes espectrogramas generadas a partir de señales de oximetría extraídas de la base de datos CHAT.

3.4.1 TRANSFER LEARNING CON RESNET50

El primer bloque experimental se llevó a cabo empleando la arquitectura ResNet50 pre-entrenada, cuyos pesos se tomaron del conjunto de datos de *ImageNet*. Como ya se ha mencionado, esta red es muy adecuada para imagen médica, ya que ha demostrado alcanzar en algunas tareas los mejores resultados en comparación de otras redes (Herrero-Tudela *et al.*, 2025).

Durante el entrenamiento se aplicó la técnica *fine-tuning*, sin congelar ninguna de las capas. Esta técnica consiste en tomar una red ya pre-entrenada con otros datos diferentes

y emplearla con tus datos permitiendo que se ajusten los parámetros al nuevo problema. Al no congelar ninguna capa, ningún parámetro queda fijo y todos son adaptables durante el entrenamiento. Está demostrado que no congelar capas es más adecuado para que la red aprenda mejor, especialmente cuando se dispone de pocos datos de entrada (Becherer *et al.*, 2019).

Inicialmente se preparan los datos para que puedan ser aceptados por la red. Dado que se utilizan como entrada los espectrogramas RGB, se redimensionan a un tamaño de 120×16.000 píxeles. Esta dimensión unifica todas las imágenes y además mantiene un equilibrio entre el elevado coste computacional y la conservación de la información de la resolución original.

A pesar de que se trabaje con un problema de regresión prediciendo el AHI, este valor se emplea posteriormente para clasificar la enfermedad en cuatro niveles de severidad. Esta es la razón por la cual se realiza un balance de los datos. Para ello, se dividen las señales atendiendo a la siguiente clasificación:

- No AOS: $AHI \leq 1$
- AOS leve: $1 \leq AHI < 5$
- AOS moderada: $5 \leq AHI < 10$
- AOS grave: $AHI \geq 10$.

Se obtiene un total de 219 imágenes correspondientes a sujetos sin AOS, 497 con AOS leve, 159 con AOS moderada y 131 con AOS severa. Dado que los grupos se encontraban desequilibrados, se aplicó una estrategia de sobremuestreo repitiendo imágenes en los grupos minoritarios con el fin de mitigar el sesgo durante el entrenamiento. Los generadores de imágenes se emplean solamente para re-escalar los píxeles de 0 a 1. Además, se usa un *batch size* de 4 debido a que el elevado tamaño de las imágenes genera limitaciones de memoria en la GPU.

En cuanto a la arquitectura del modelo, ResNet50 se usa como *backbone* pero se eliminan sus últimas capas para añadir una cabeza personalizada al estudio añadiendo capas *Dropout*, *GlobalAveragePooling* y *Dense*.

Por una parte, se añade un *GlobalAveragePooling2D* que promedia todos los valores en cada canal para reducir la dimensionalidad, resultando en un vector 1×1×3 y minimizando así el sobreajuste. También se aplican tres capas de *Dropout* con tasa constante en todas (Jiménez-García *et al.*, 2024). Esta técnica consiste en desactivar un porcentaje de neuronas en cada *batch* para que la red no aprenda siempre de los mismos patrones de activación tome nuevos caminos y aumente la generalización (Moradi, Berangi and Minaei, 2020). En este caso se emplea *Dropout* de 0.2, ya que es adecuado en CNN cuando se trata con la señal de SpO_2 para diagnóstico de apnea (Tyagi and Agarwal, 2023). Estas capas se intercalan con tres capas *Dense* con su función de activación *ReLu*, de las cuales, la última cuenta con una única neurona y función de activación lineal para adecuar la salida al problema de regresión, es decir, para poder predecir un número continuo.

Además del *Dropout* y el balanceo de los datos, se aplican más técnicas de regularización. Estas técnicas se añaden para evitar que la red se ajuste demasiado a los datos de entrenamiento y el modelo tenga robustez para funcionar bien ante datos nuevos (Moradi, Berangi and Minaei, 2020). Así, se ha aplicado regularización *L2*, que penaliza los pesos cuando aumentan demasiado, normalmente a causa del ruido. Asimismo, se ha empleado *Early stopping*, configurado con una paciencia de 40 épocas, con el objetivo de detener el entrenamiento si no se observa mejora en la pérdida de validación durante dicho intervalo (Moradi, Berangi and Minaei, 2020).

En las redes se realiza una optimización para ir ajustando los parámetros del modelo mientras se minimiza una función de pérdidas (Sun, 2020). Existen diferentes métodos para llevar a cabo esto, pero el optimizador seleccionado fue Adam (Herrero-Tudela *et al.*, 2025). Por otra parte, el *learnig rate* controla los pasos que da el optimizador cuando ajusta los parámetros. Por esta razón, también se aplican métodos adaptativos que ajustan esta tasa según la evolución del modelo, permitiendo una convergencia más rápida y estable (Sun, 2020). Para llevarlo a cabo se ha implementado *ReaduceLROnPLateau*, que reduce el *learning rate* a la mitad hasta un valor máximo de 10^{-8} si la pérdida de validación no mejora en 5 épocas.

En cuanto a la función de pérdida, tras un proceso comparativo, se seleccionó *mean squared logarithmic error*. Esta función es adecuada debido a que la variable AHI empleada en este trabajo toma valores positivos. Además, penaliza en mayor medida las subestimaciones, lo cual es importante en el ámbito clínico (Jadon, Patil and Jadon, 2022).

En la Figura 6 se presenta en forma de esquema el proceso llevado a cabo en este trabajo. Una vez obtenidos los espectrogramas de la señal de SpO2, estos se emplean como entrada a la arquitectura ResNet50 para la estimación del AHI y la posterior clasificación de la AOS.

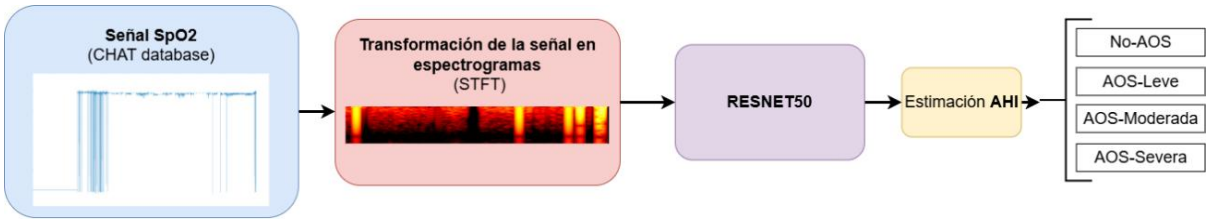

Figura 6. Esquema del procedimiento empleado en el estudio desde la señal hasta la estadificación de la enfermedad.

3.4.2 TRANSFER LEARNING CON RESNET PRENTRENADA CON IMÁGENES DE RETINA

Tras completar el entrenamiento utilizando ResNet50 con pesos pre-entrenados en el conjunto de datos *ImageNet*, se llevó a cabo un segundo bloque experimental basado en el mismo *backbone*. En esta ocasión, se utilizó una versión de ResNet50 pre-entrenada con imágenes de retina, con el objetivo de evaluar si la similitud cromática entre las

retinografías, predominantemente compuestas por tonos rojizos, y los espectrogramas empleados en este estudio podría aportar ventajas en el proceso de aprendizaje automático. En el estudio de Herrero-Tudela et al., (2025), que propone un modelo de DL junto a técnicas de SHAP para clasificar la retinopatía diabética con imágenes de este tipo, se alcanza el rendimiento más elevado con ResNet50 comparado con otros 6 tipos de arquitecturas diferentes.

Se procede de la misma forma y con los mismos hiperparámetros que en las pruebas anteriores, salvo en el proceso de creación del modelo. En este caso, se inspecciona el nuevo modelo para comprobar el lugar donde finaliza el *backbone*. Esto permite saber qué capas pertenecen a la cabeza del modelo con el fin de no incluirlas.

Dado que las imágenes utilizadas para el entrenamiento del modelo de retina presentan una resolución significativamente inferior a la de los espectrogramas empleados en este estudio, fue necesario adaptar la red a un nuevo tamaño de entrada. Para ello, se definió una nueva capa de entrada con dimensiones de 120×16.000 píxeles, que reemplazó a la capa original de entrada del modelo pre-entrenado. Tras este paso, se construye un nuevo *backbone* copiando una a una las capas del modelo, tanto en la configuración como en los pesos. Por último, se añaden las mismas últimas capas que en el modelo entrenado anteriormente, con el fin de mantener la coherencia metodológica y facilitar una comparación objetiva entre ambos enfoques.

3.4.3 IMPLEMENTACIÓN DE SHAP

Una vez identificado el modelo óptimo entre todos los entrenados, se aplica el paquete *shap.Explainer* para calcular los valores SHAP correspondientes a las imágenes y así determinar las regiones que han tenido mayor influencia en las predicciones del modelo. En teoría SHAP asigna valores a cada píxel, pero debido al coste computacional, el método implementado agrupa automáticamente regiones de la imagen en superpíxeles. Cada superpíxel se considera como una unidad en el análisis de contribución, por lo que los valores SHAP reflejan la importancia de regiones de la imagen en lugar de píxeles individuales.

Para la representación gráfica de los valores SHAP, Se ha empleado el mapa de color '*bwr'* que emplea el rojo para zonas de la imagen que influyen en la predicción, blanco para el neutro y azul para las que disminuyen la contribución (Herrero-Tudela *et al.*, 2025).

Como se trata de imágenes de alta longitud en píxeles, también se desarrolla un código para superponer por separado el mapa SHAP sobre el espectrograma en su tamaño original y en escala de grises. Para ello, se diseña un mapa de color personalizado, pero que mantenga los mismos colores que el anterior, el color azul al valor SHAP negativo, el rojo al positivo y gris al valor 0. El objetivo es poder ver localmente lo que ocurre en cada zona de la imagen. Además, la escala de grises permite que los colores del mapa '*Hot'* no se confundan con los del mapa.

Durante la experimentación, se analizó el parámetro *max_evals*, que determina el número máximo de evaluaciones del modelo durante el análisis SHAP. Se inició con un valor de 1.000 y, tras pruebas incrementales, se estableció un valor final de 8.000 evaluaciones. Cuanto mayor es el número de evaluaciones, más detalladas son las explicaciones de los mapas, pero mayor es el tiempo computacional. El número final establecido de 8.000 aporta un equilibrio adecuado entre la calidad de la explicación y el coste computacional.

La forma en la que se enmascara las características no siempre es la misma. En este caso, tras una fase de experimentación, se decidió emplear el *masker inpaint-telea.* Este algoritmo rellena las características ocultas mediante el método *Telea*, que utiliza la información de los pixeles de la propia imagen que están alrededor de estas regiones, haciendo el análisis más coherente (Telea, 2004).

3.6 ANÁLISIS ESTADÍSTICO

En el presente trabajo, se estima el AHI mediante un enfoque de regresión. Tras esto, se emplea este valor continuo en la clasificación de la enfermedad en cuatro severidades. Con el fin de evaluar y comparar el rendimiento de los modelos en función de esta clasificación, se emplea la matriz de confusión junto con algunas métricas derivadas de ella.

La matriz de confusión es una herramienta que se emplea tanto en problemas binarios como en multiclase. Su estructura permite comparar las etiquetas reales (organizadas por filas) con las etiquetas predichas por el modelo (organizadas por columnas), haciendo que se generen los siguientes elementos (Grandini, Bagli and Visani, 2020):

- Verdaderos positivos (VP): sujetos de la clase positiva correctamente clasificados.
- Verdaderos negativos (VN): sujetos de la clase negativa correctamente clasificados.
- Falsos positivos (FP): sujetos de la clase negativa clasificados como positivos.
- Falsos negativos (FN): sujetos de la clase positiva clasificados como negativos.

En el contexto de clasificación multiclase adoptado en este trabajo, la matriz de confusión no solo permite examinar el rendimiento específico de cada clase de apnea, sabiendo que cada posición (i, j) de ésta representa los sujetos de la clase i clasificados como clase j, sino que también permite dar un rendimiento 4-clases teniendo en cuenta todos los aciertos y todos los fallos cometidos en las predicciones y aportando una visión más global (Grandini, Bagli and Visani, 2020).

3.5.1 MÉTRICAS DE RENDIMIENTO DE LA ESTIMACIÓN DEL AHI

Puesto que en este trabajo se realiza inicialmente una regresión para estimar el índice AHI, es importante establecer unas métricas que permitan cuantificar la proximidad entre

los valores predichos y los valores reales. Para ello, se hace uso del ICC y del diagrama de Bland-Altman.

El ICC mide el grado de fiabilidad teniendo en cuenta la variabilidad entre los sujetos debido al error. Existen diferentes formas de cálculo, pero en este trabajo se emplea el ICC (2,1). Esta forma es adecuada para la medición de una misma variable por sujeto y además permite generalizar los resultados. Dicha métrica se calcula como se indica en la Ecuación 6 (Weir, 2005).

$$ICC(2,1) = \frac{MSS - MSE}{MSS + (k-1)\cdot MSE + \frac{k\cdot(MST - MSE)}{n}} \quad (6)$$

donde MSS representa la variabilidad entre los sujetos, MST la variabilidad entre las mediciones, MSE el error cuadrático medio, k el número de mediciones por sujeto y *n* el número de sujetos.

Se pueden establecer unos rangos de manera que ICC por encima de 0.90 indica fiabilidad excelente, ICC entre 0.75-09 buena, ICC entre 0.5-0.75 moderada e ICC por debajo 0.5 pobre (Koo and Li, 2016).

También se hace uso del diagrama de Bland-Altman, que permite evaluarla concordancia entre los valores de AHI estimados por la red y los reales calculando las diferencias individuales de cada par de mediciones. Esto permite descubrir si existe sesgo sistemático y si el modelo subestima o infraestima el AHI respecto a los valores reales (Giavarina, 2015). En esta representación, en el eje x se plasma el promedio de las medidas y en el eje y la diferencia entre estas. También se indica mediante líneas el sesgo promedio y los límites de acuerdo calculados como ± 1.96 veces la desviación estándar de la diferencia. Para considerar que un diagrama Bland-Altman está representando unos buenos resultados, la mayor parte de las mediciones (alrededor del 95%) tienen que entrar entre los límites de acuerdo superior e inferior. Además, el sesgo promedio, debe situarse cerca del 0, lo que significaría ausencia de sobreestimación o infraestimación sistemática (Giavarina, 2015).

3.5.2 MÉTRICAS DE RENDIMIENTO DE LA CLASIFICACIÓN

Para la evaluación comparativa del rendimiento de los distintos modelos de clasificación, se emplean diferentes métricas. En primer lugar, el coeficiente *kappa de Cohen* es una métrica particularmente útil en escenarios con clases desbalanceadas (Grandini, Bagli and Visani, 2020). *Kappa de Cohen* mide la concordancia entre las clases estimadas y las etiquetas teniendo en cuenta los acuerdos realizados por el azar. Su cálculo se expresa en la Ecuación 7 (Grandini, Bagli and Visani, 2020).

$$kappa = \frac{P_o - P_e}{1 - P_e} \quad (7)$$

donde Po es la proporción de concordancia observada y Pe la proporción de concordancia debida al azar. El coeficiente kappa toma valor 1 cuando hay acuerdo perfecto, 0 cuando la concordancia es igual a la esperada por azar, y valores negativos cuando el acuerdo es inferior al esperado aleatoriamente,(Grandini, Bagli and Visani, 2020).

Una vez seleccionado el modelo con mejor rendimiento global, se procede a una evaluación más específica mediante el cálculo de diversas métricas para cada una de las categorías de severidad de la apnea. Estas métricas incluyen: precisión, sensibilidad, especificidad, valor predictivo negativo, valor predictivo positivo, *likelihood ratio* positivo y *likelihood ratio* negativo.

La métrica precisión (*Accuracy, Acc*) sirve para medir el grado de acierto del modelo en las estimaciones proporcionadas. Este valor refleja los sujetos que han sido bien clasificados, es decir, los aciertos que se observan en la diagonal de la matriz. Para un problema multiclase, el *Acc* para cada clase se calcula como se muestra en la Ecuación 8 (Grandini, Bagli and Visani, 2020):

$$Acc = \frac{VP + VN}{VP + VN + FP + FN} \tag{8}$$

También se calcula la sensibilidad (Se), que representa la capacidad del modelo para acertar los casos positivos reales. Para cada una de las clases se calcula como se muestra en la Ecuación 9 (Vihinen, 2012):

$$Se = \frac{VP}{VP + FN} \tag{9}$$

Por otra parte, la especificidad (Sp) es una métrica que refleja la capacidad para acertar los casos negativos reales. Para cada una de las clases se calcula como se muestra en la Ecuación 10 (Vihinen, 2012):

$$Sp = \frac{VN}{VN + FP} \tag{10}$$

Además, se hace uso del valor predictivo positivo (VPP) que indica la probabilidad que existe de que un caso clasificado como positivo sea realmente positivo. Su cálculo se refleja en la Ecuación 11 (Vihinen, 2012):

$$VPP = \frac{VP}{VP + FP} \tag{11}$$

De una forma similar se emplea el valor predictivo negativo (VPN) que, en este caso, es la probabilidad de que un caso clasificado como negativo sea realmente negativo. Su cálculo se refleja en la Ecuación 12 (Vihinen, 2012):

$$VPN = \frac{VN}{VN + FN} \tag{12}$$

Por último, se calcula *likelihood ratio* positivo y negativo. El *likelihood ratio* positivo (LR+) indica cuán más probable es clasificar como positivo a un paciente que es realmente positivo en comparación con uno que es negativo. Esto se calcula como se muestra en la Ecuación 13 (Roman Jaeschke, Gordon H. Guyatt and David L. Sackett, 1994):

$$LR+ = \frac{Se}{1 - Sp} \tag{13}$$

Por el contrario, el *likelihood ratio* negativo (LR-) indica cuán más probable es clasificar como negativo a un paciente que es realmente positivo comparación con uno que es negativo. Esto se calcula como se muestra en la Ecuación 12 (Roman Jaeschke, Gordon H. Guyatt and David L. Sackett, 1994):

$$LR- = \frac{1 - Se}{Sp} \tag{14}$$

Un LR+ alto indica que el modelo es bueno para confirmar la enfermedad y un LR- bajo indica que el modelo es bueno para descartar la enfermedad. De esta forma, se consideran los siguientes rangos para descartar o confirmar la enfermedad (Roman Jaeschke, Gordon H. Guyatt and David L. Sackett, 1994):

- *LR*+>10 y *LR*-<0.1: evidencia fuerte.
- *LR*+: 5-10 y *LR*-: 0.1-0.2 evidencia moderada.
- *LR*+: 2-5 y *LR*-: 0.2-0.5 evidencia pequeña.
- *LR*+: 1-2 y *LR*-: 0.5-1 evidencia con escasa relevancia.

CAPÍTULO 4: RESULTADOS

4.1 OPTIMIZACIÓN DE HIPERPARÁMETROS

Con el objetivo de mejorar los resultados y adaptar los modelos al problema particular de este estudio, se llevó a cabo un proceso de optimización de hiperparámetros en el grupo de validación. Esta optimización se centró solamente en el número de neuronas de las dos primeras capas *Dense* añadidas a la cabeza. La tercera no se modifica debido a que obligatoriamente debe tener una neurona para adaptarse a la salida del problema de regresión.

Se varía el número de la primera *Dense* entre 2048, 1024 y 512 neuronas, mientras que la segunda entre 512 y 256. Se evaluaron cinco configuraciones distintas, restringiendo que el número de neuronas en la primera capa fuera mayor que en la segunda:

- (2048, 512)
- (1024, 512)
- (2048, 256)
- (1024, 256)
- (512, 256)

Cada una de estas combinaciones fue entrenada tanto con la arquitectura ResNet50 pre-entrenada con pesos de *ImageNet* como con una versión pre-entrenada utilizando imágenes de retina con el fin de elegir el mejor modelo. El coeficiente *kappa* de 4 clases para cada combinación del hiperparámetros en ambas redes sobre el conjunto de validación se recoge en las Tablas 3 y 4.

ResNet50	
Nº neuronas (1º *Dense*, 2º *Dense*)	**kappa**
(2048, 512)	0.499
(1024, 512)	**0.554**
(2048, 256)	0.520
(1024, 256)	0.549
(512, 256)	0.517

Tabla 3. Coeficiente kappa para cada combinación del número de neuronas de las capas Dense en ResNet50.

ResNet50 + Retina	
Nº neuronas (1º *Dense*, 2º *Dense*)	**kappa**
(2048, 512)	0.539
(1024, 512)	0.521
(2048, 256)	0.537
(1024, 256)	**0.543**
(512, 256)	0.540

Tabla 4. Coeficiente kappa para cada combinación del número neuronas de las capas Dense en ResNet50 pre-entrenada con imágenes de retina.

4.2 RESULTADOS DE LOS MEJORES MODELOS

Tras el proceso de optimización, se seleccionan los modelos cuyos hiperparámetros han dado un valor de *kappa* más elevado en el grupo de validación, ya que esto indica un mayor grado de concordancia entre las predicciones y las etiquetas reales.

Con este criterio, el modelo que alcanzó el mejor valor para ResNet50 fue la configuración de 1024 neuronas en la primera capa *Dense* y 512 en la segunda, con un coeficiente *kappa* de 0.533 en el grupo de test. La matriz de confusión derivada de la clasificación de la apnea a partir del AHI se muestra en la Figura 7. se presenta en la Tabla 5, las métricas de rendimiento de la clasificación binaria de la AOS infantil por parte del modelo ResNet50, atendiendo a los umbrales de AHI 1,5 y 10 e/h. En la Tabla 6 se recoge el *ICC*, el *Acc* 4-clases y el *kappa* 4-clases para el modelo ResNet50. Para completar los resultados se realiza un diagrama de *Bland-Altman* (Figura 8) donde el eje x representa el promedio entre AHI real y el estimado y el eje y, la diferencia entre estos mismos. La línea central roja es el sesgo promedio y las líneas superiores e inferiores verdes son los límites de acuerdo.

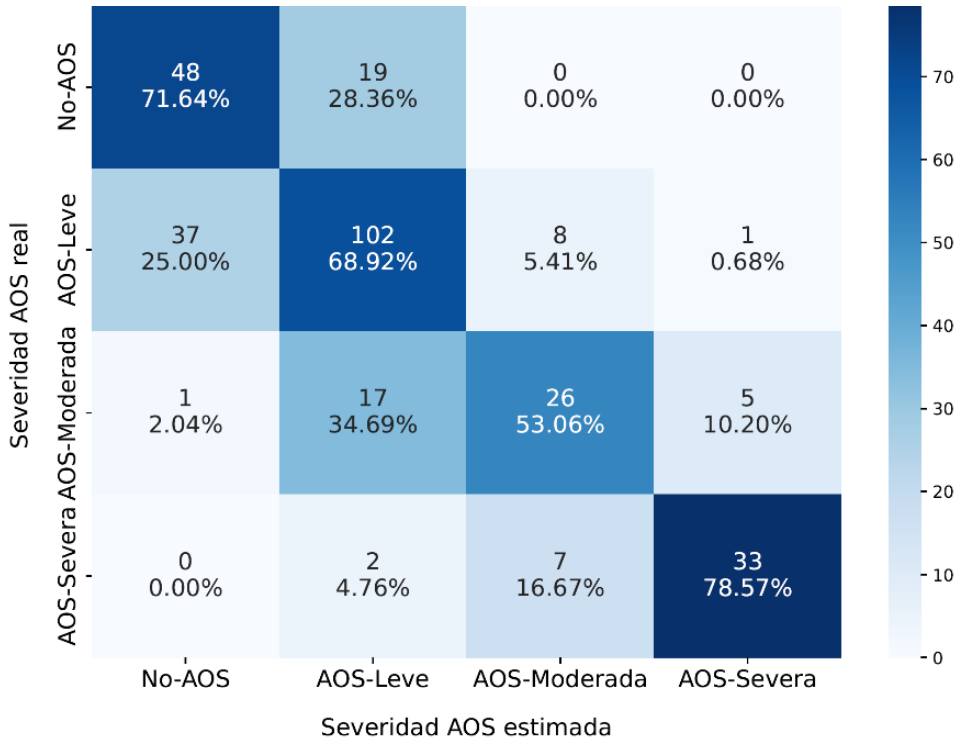

Figura 7. Matriz de confusión del grupo de test obtenida del mejor modelo Resnet50 entrenada con espectrogramas de la señal SpO_2 de la base de datos CHAT.

ResNet50							
	Acc	***Se***	***Sp***	***VPP***	***VPN***	***LR+***	***LR-***
$1\ e/h$	81.37%	84.10%	71.64%	91.36%	55.81%	2.97	0.22
$5\ e/h$	90.52%	78.02%	95.81%	88.75%	91.15%	18.64	0.23
$10e/h$	95.10%	78.57%	97.73%	84.62%	96.63%	34.57	0.22

Tabla 5. Métricas calculadas para el mejor modelo ResNet50para cada umbral.

ResNet50	
ICC	0.891
Acc 4-clases	68.30%
Kappa 4-clases	0.5326

Tabla 6. ICC para el mejor modelo ResNet50.

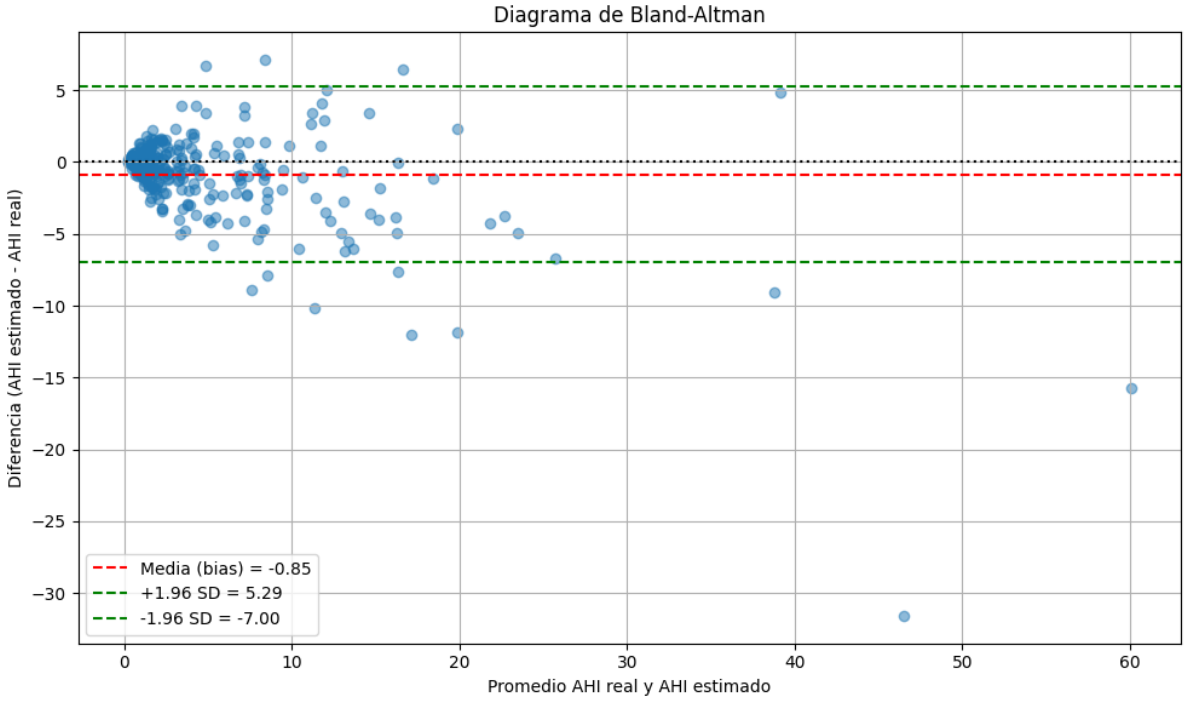

Figura 8. Diagrama Bland-Altman para modelo el mejor modelo ResNet50.

De la misma forma, se elige el mejor modelo para la red ResNet50 pre-entrenada con imágenes de retina. La combinación de neuronas óptima fue 1024 y 256 para la primera y segunda *Dense* respectivamente, con un coeficiente *kappa* en el grupo de test de 0.489. Al igual que para la red anterior, para recoger los resultados se emplean la matriz de

confusión (Figura 9), tabla de métricas (Tablas 7 y 8) y diagrama de *Bland-Altman* (Figura 10)

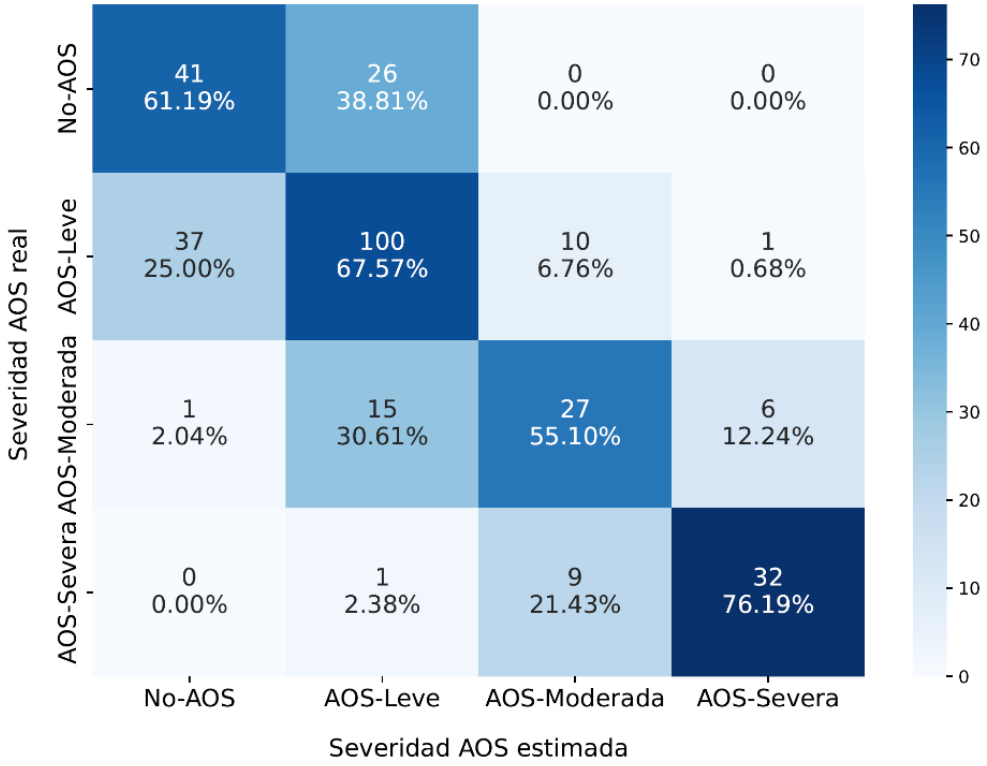

Figura 9. Matriz de confusión obtenida del mejor modelo Resnet50 + Retina entrenada con espectrogramas de la señal SpO_2 de la base de datos CHAT.

ResNet50 + Retina							
	Acc	***Se***	***Sp***	***VPP***	***VPN***	***LR+***	***LR-***
$1\ e/h$	79.08%	84.10%	61.19%	88.55%	51.90%	2.17	0.26
$5\ e/h$	90.85%	81.32%	94.88%	87.06%	92.31%	15.89	0.20
$10\ e/h$	94.44%	76.19%	97.35%	82.05%	96.25%	28.73	0.24

Tabla 7. Métricas calculadas para el mejor modelo ResNet50 + Retina para cada umbral.

ResNet50 + Retina	
ICC	0.911
Acc - 4 clases	65.36%
Kappa – 4 clases	0.4887

Tabla 8. ICC para el mejor modelo ResNet50 + Retina.

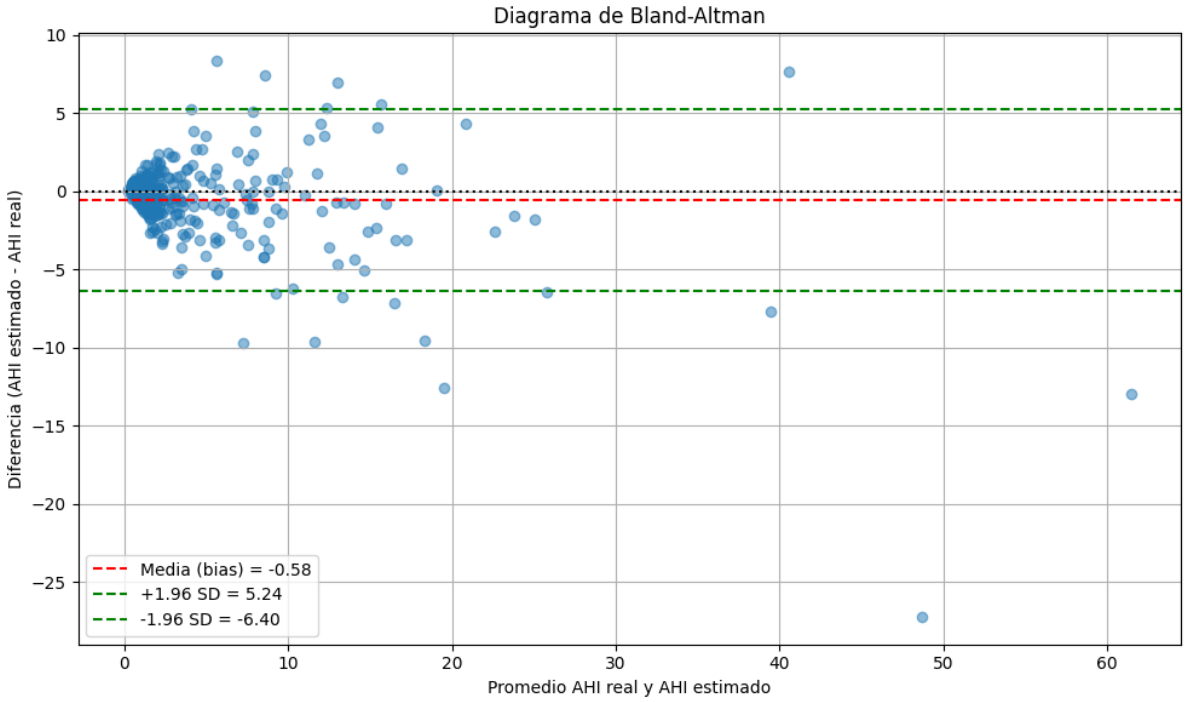

Figura 10.Diagrama Bland-Altman para modelo el mejor modelo ResNet50 + Retina.

4.3 RESULTADOS DE SHAP

A partir de la comparación entre ambos modelos, se seleccionó el modelo ResNet50 pre-entrenado con *ImageNet* (configuración 1024-512) como el más idóneo para aplicar técnicas de interpretación con SHAP, al haber obtenido un mayor coeficiente *kappa*. Se seleccionaron dos sujetos por clase de severidad de apnea: uno con la mejor estimación (menor error absoluto entre AHI estimado y real) y otro con la peor. Los resultados se resumen en la Tabla 9.

	AHI real	**AHI estimado**	**Error absoluto (|AHI real – AHI estimado|)**	**Estimación (Mejor/Peor)**
$AHI \leq 1$	0.533	0.538	0.005	Mejor
	0.606	2.802	2.196	Peor
$1 \leq AHI < 5$	1.328	1.347	0.019	Mejor
	4.855	11.955	7.101	Peor
$5 \leq AHI < 10$	8.182	8.050	0.132	Mejor
	8.213	2.445	5.768	Peor
$AHI \geq 10$	16.396	16.347	0.049	Mejor
	62.313	30.741	31.572	Peor

Tabla 9. Mejores y peores estimaciones del AHI para cada clase de apnea.

Las explicaciones SHAP se visualizan mediante la superposición del mapa de calor sobre la señal de SpO_2 original, incluyendo también las anotaciones de eventos apnéicos y su espectrograma asociado. En las Figuras 11 y 12 se presentan las explicaciones para los sujetos de la clase No-AOS con la mejor y la peor estimación del índice AHI,

respectivamente. Las Figuras 13 y 14 corresponden a los sujetos de la clase AOS leve, también diferenciando entre la predicción más precisa y la de mayor error. De igual modo, las Figuras 15 y 16 muestran los resultados para la clase AOS moderada, y las Figuras 17 y 18 recogen las explicaciones generadas para la clase AOS severa. Estas visualizaciones permiten interpretar de manera localizada qué regiones del espectrograma han sido más determinantes en la predicción del modelo, lo que contribuye significativamente a aumentar la transparencia del proceso diagnóstico automático basado en DL.

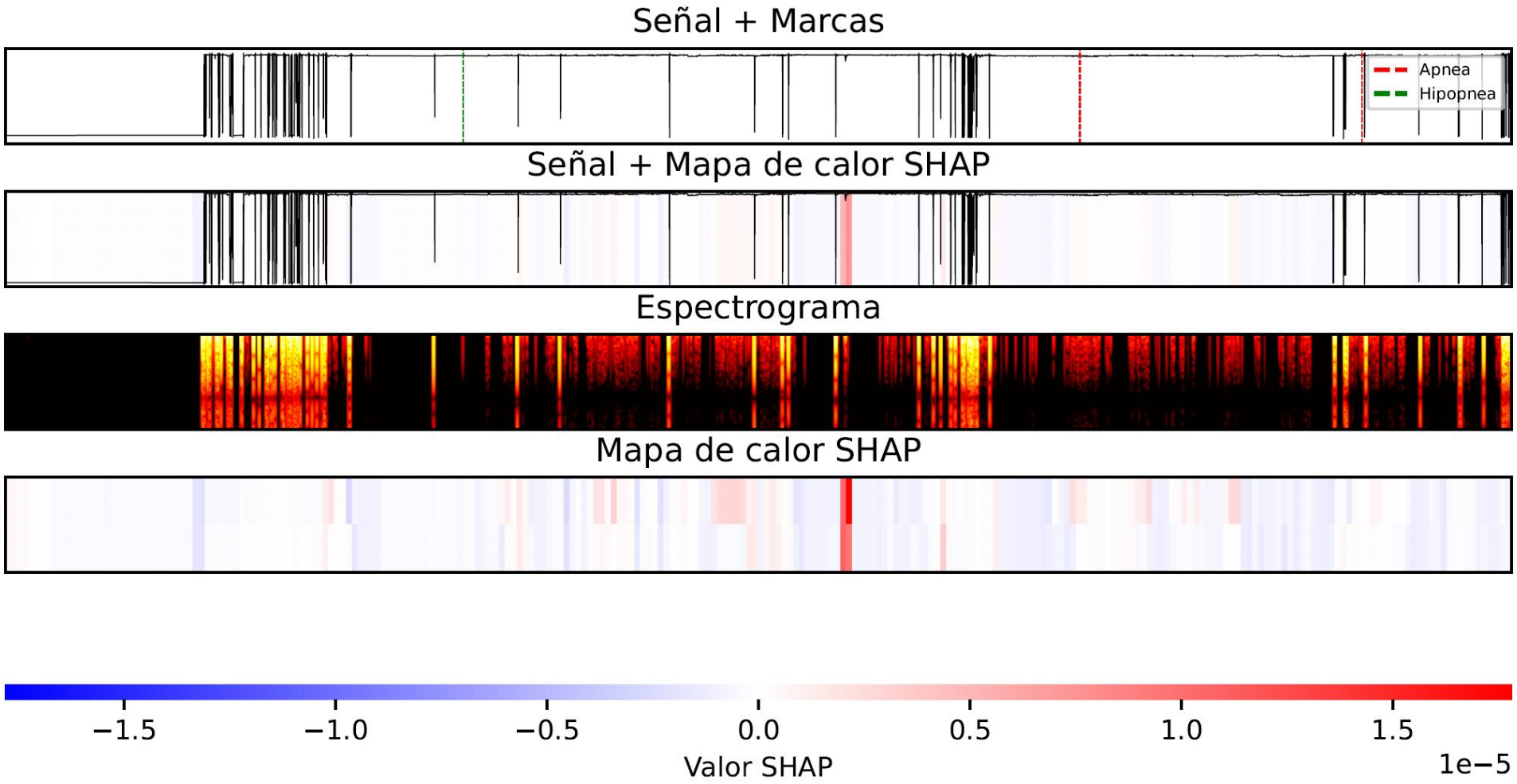

Figura 11. Análisis SHAP para el sujeto de clase No-AOS con la mejor estimación del AHI.

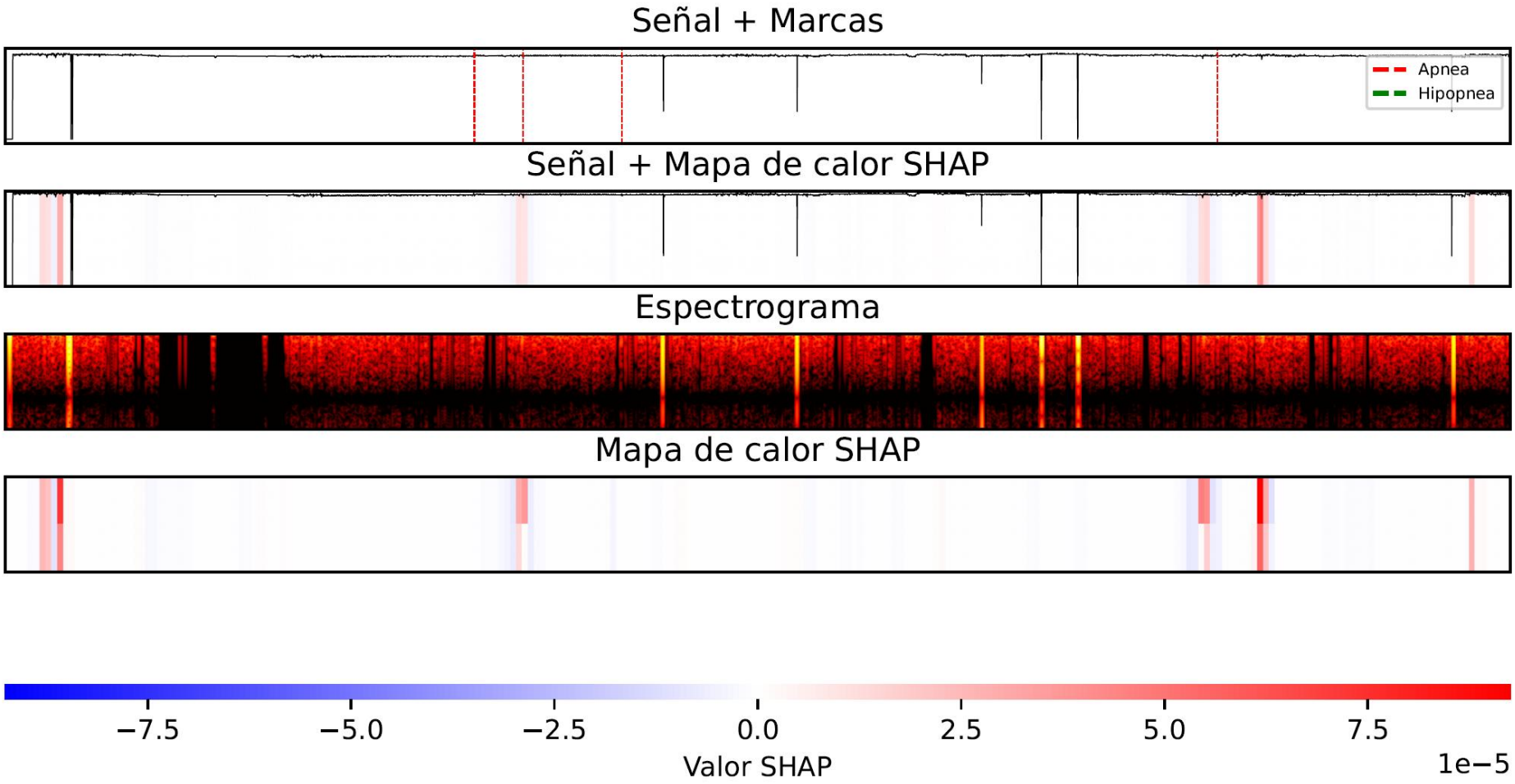

Figura 12. Análisis SHAP para el sujeto de clase No-AOS con la peor estimación del AHI.

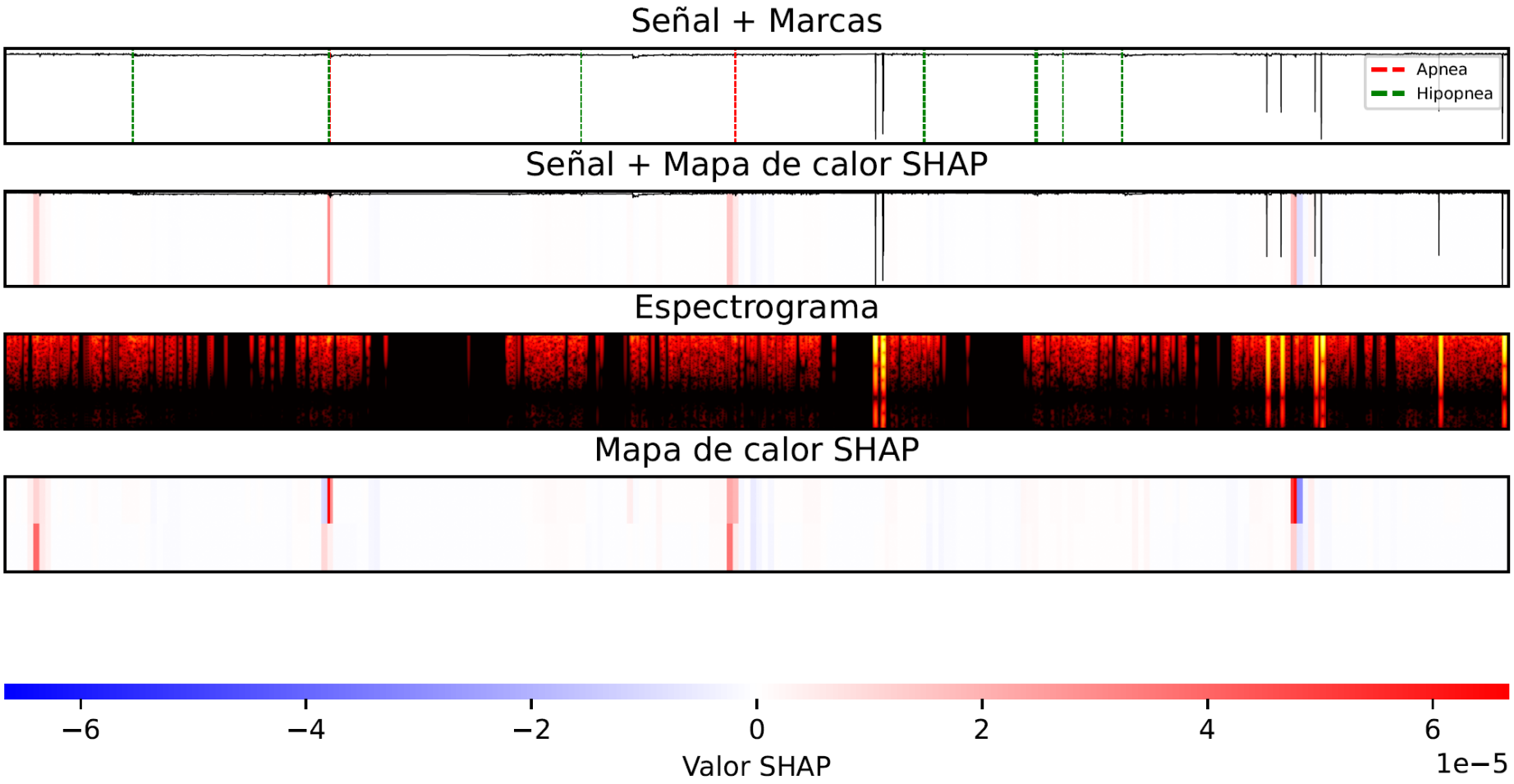

Figura 13. Análisis SHAP para el sujeto de clase AOS leve con la mejor estimación del AHI.

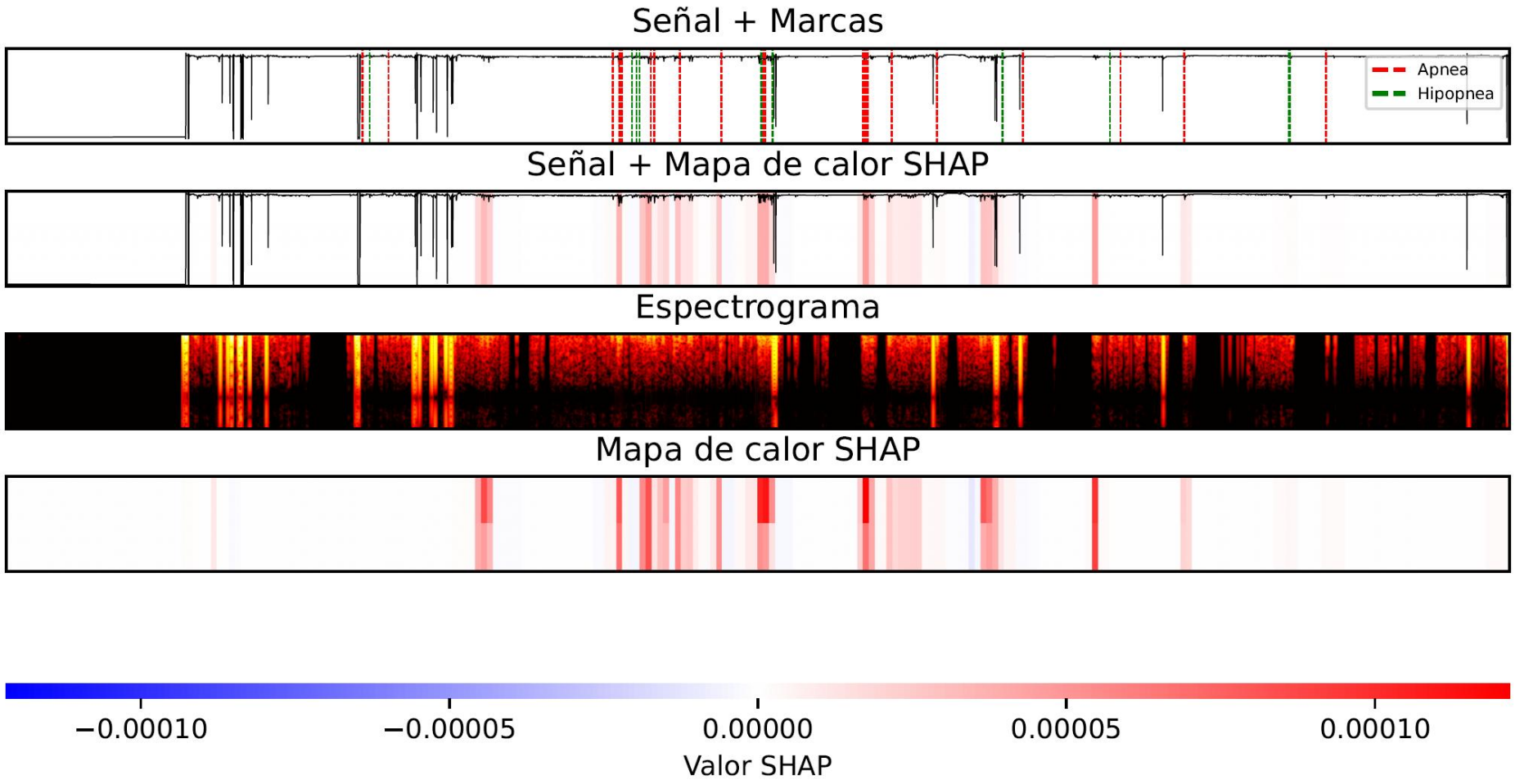

Figura 14. Análisis SHAP para el sujeto de clase AOS leve con la peor estimación del AHI.

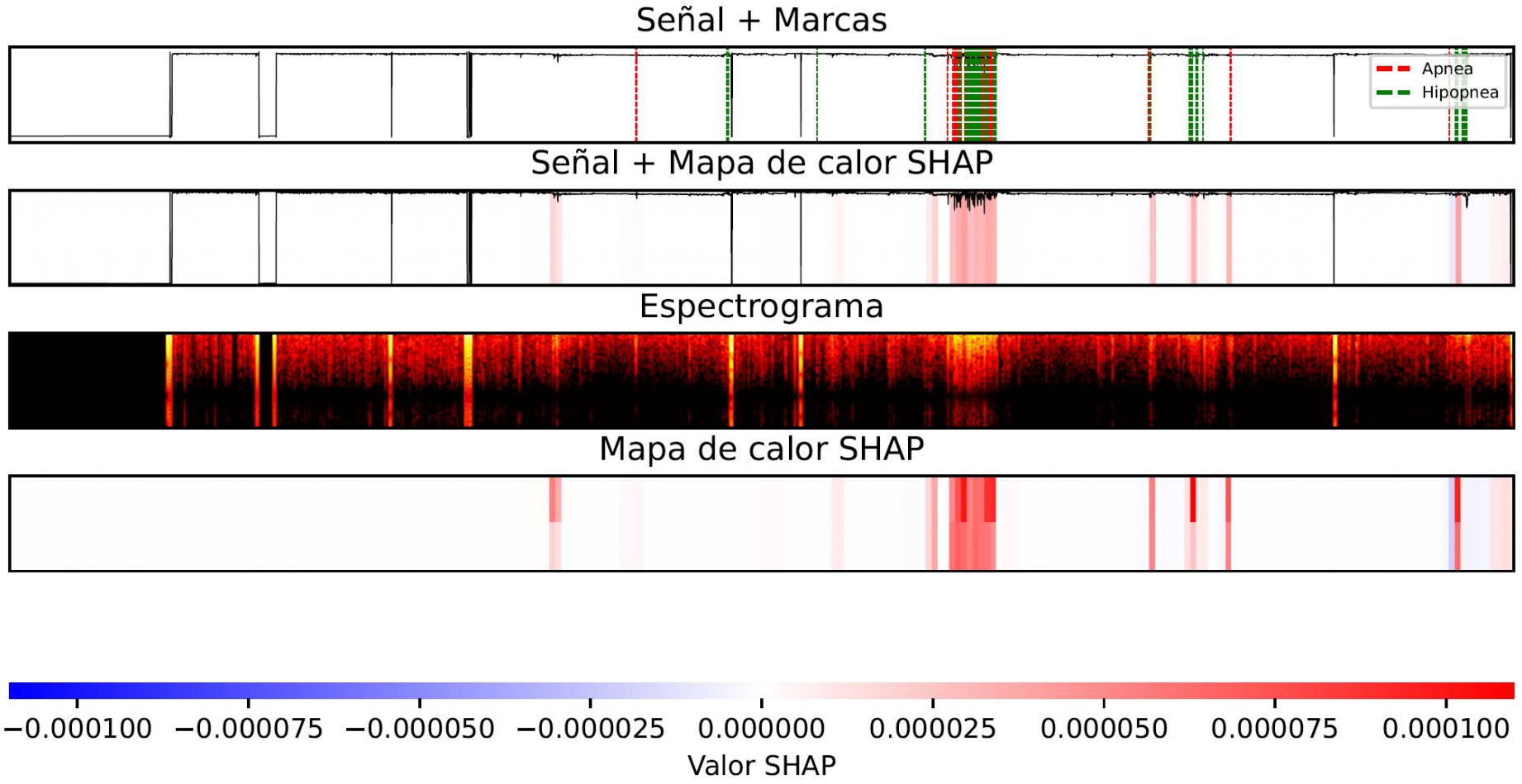

Figura 15. Análisis SHAP para el sujeto de clase AOS moderada con la mejor estimación del AHI.

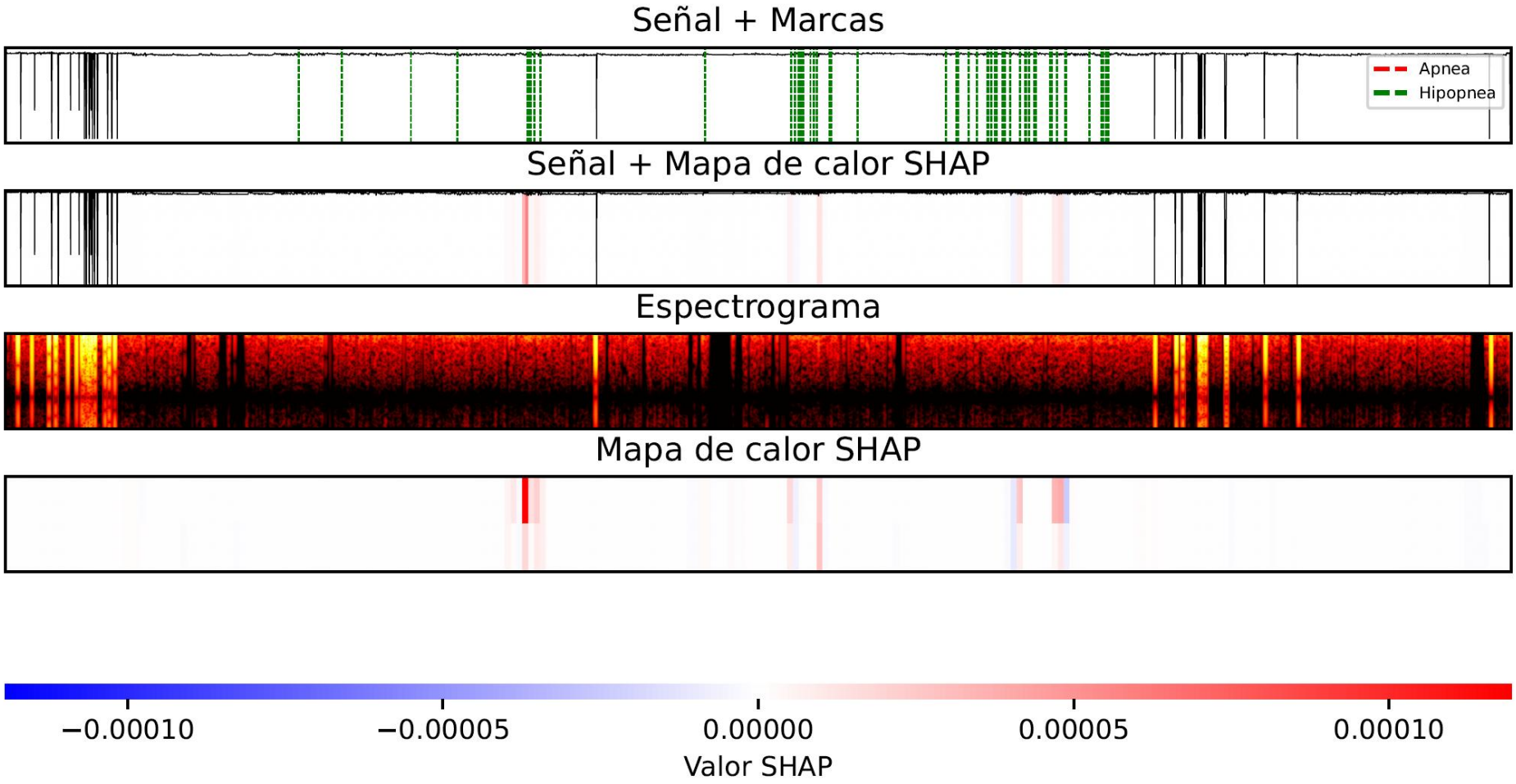

Figura 16. Análisis SHAP para el sujeto de clase AOS moderada con la peor estimación del AHI.

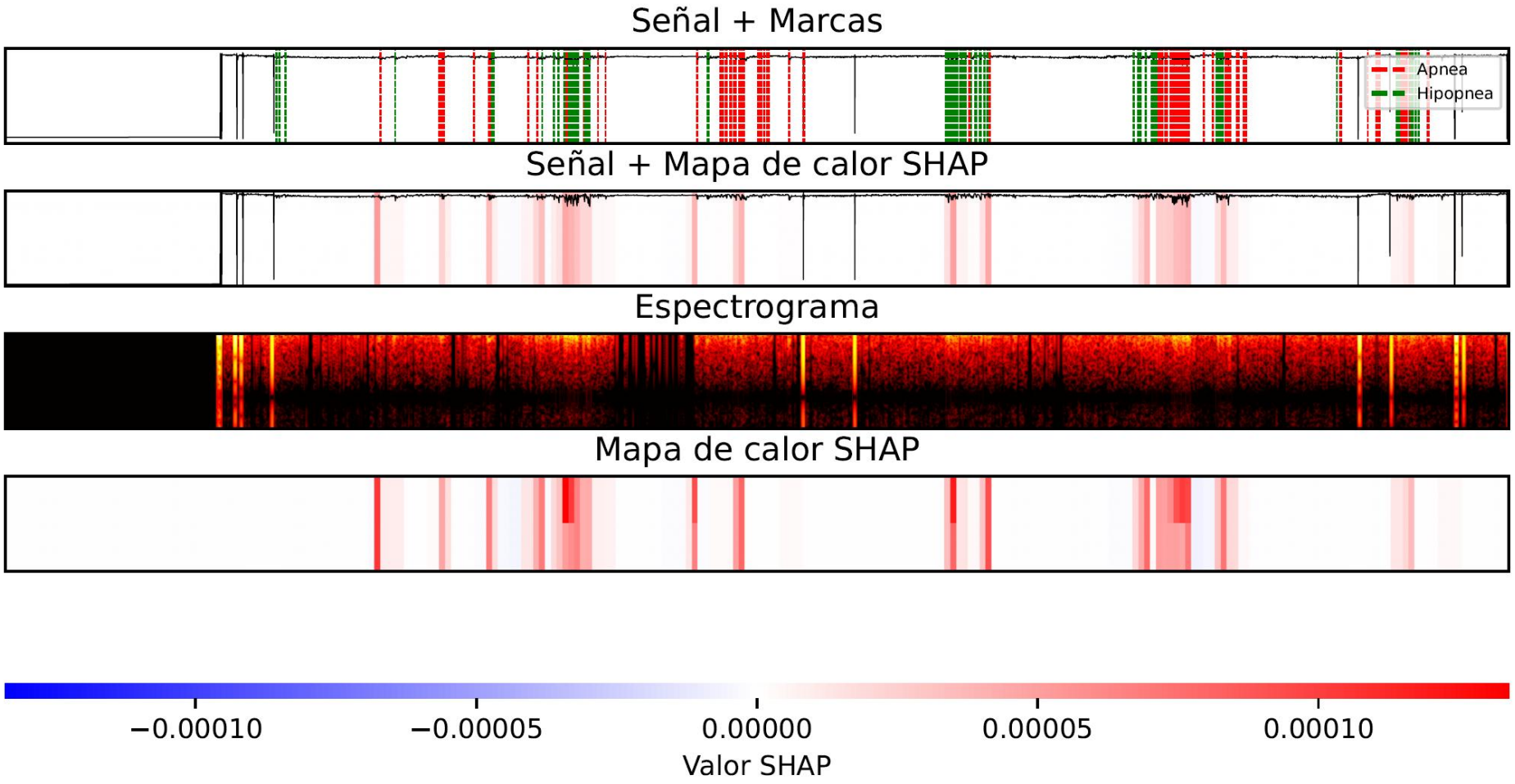

Figura 17. Análisis SHAP para el sujeto de clase AOS severa con la mejor estimación del AHI.

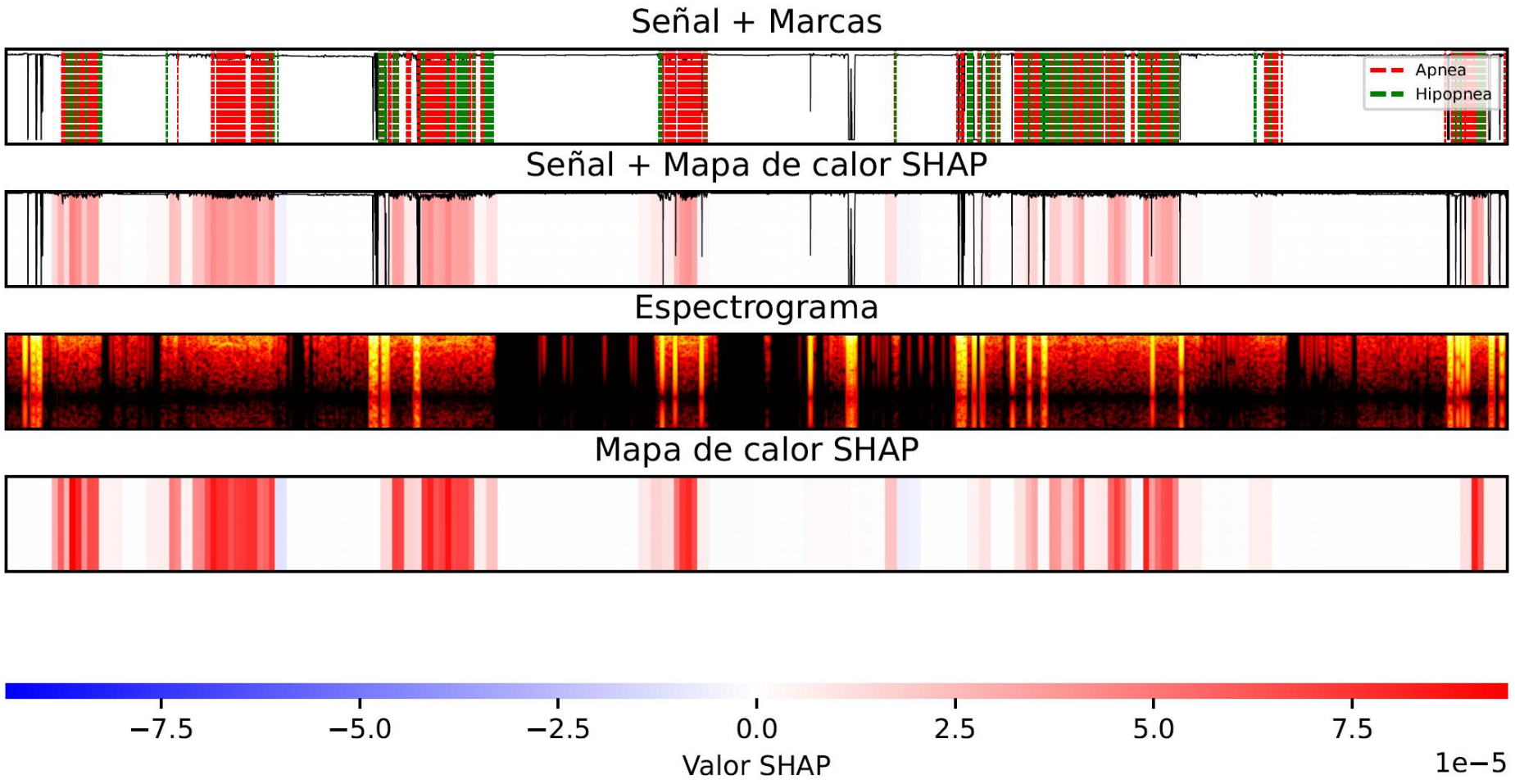

Figura 18. Análisis SHAP para el sujeto de clase AOS severa con la peor estimación del AHI.

CAPÍTULO 5: DISCUSIÓN DE LOS RESULTADOS

5.1 CLASIFICACIÓN DE LA SEVERIDAD DE LA APNEA

Una vez presentados los resultados de los modelos ResNet50 y ResNet50 pre-entrenada con imágenes de retina (ResNet50 + Retina), se procede a discutir sus aspectos más importantes en la estadificación de la AOS infantil. Para ello, en este apartado se atiende a los datos anteriormente presentados en las Tablas 5 y 7.

Para el umbral 10 e/h, ambos modelos obtienen sus mejores resultados en cuanto a precisión. En el caso de ResNet50, con un *Acc* del 95.10% y ResNet50+Retina del 94.44%. Además de la precisión, ResNet50 también alcanza resultados ligeramente superiores que ResNet50 + Retina en el resto de las métricas como *Se* (78.57% frente 76.19%), *Sp* (97.73% frente 97.35%), *VPP* (84.62% frente 82.05%), *VPN* (96.63% frente 96.25%) *LR*+ (34.57 frente 28.73) y *LR*- (0.22 frente 0.24). Destaca el *LR*+ del modelo ResNet50 con 34.6123. Esto significa que un resultado positivo en este modelo se asocia con una probabilidad aproximadamente 35 veces mayor de padecer AOS severa, en comparación con un paciente con resultado negativo. El rendimiento superior en esta clase puede deberse a que los eventos apneicos son más marcados y fáciles de distinguir al tratarse del caso severo.

Para el umbral 5 e/h la mejoría de resultados de un modelo en comparación con el otro es menos evidente. ResNet50 supera a ResNet50 + Retina en las métricas *Sp* (95.81% frente 94.88%), *VPP* (84.75% frente 87.06%) y *LR*+ (18.64 frente 15.89), lo que implica mayor robustez para minimizar los falsos positivos. Sin embargo, ResNet50 es inferior a ResNet50 + Retina en el resto de las métricas: *Acc* (90.52% frente 90.85%), *Se* (78.02% frente 81.32%), *VPN* (91.15% frente 92.31%) y *LR*- (0.23 frente 0.20). Esto significa que, para este umbral, ResNet50 + Retina es mejor para la minimización de falsos negativos.

Por último, para el umbral 1 e/h, al igual que para el umbral 10 e/h, se mantiene el carácter superior de ResNet50 en todas las métricas: *Acc* (81.37% frente 79.08%), *Sp* (71.64% frente 61.19%), *VPP* (91.36% frente 88.55%), *VPN* (55.81% frente 51.90%), *LR*+ (2.97 frente 2.17) y *LR*- (0.22 frente 0.26). Sin embargo, la sensibilidad es igual para ambos modelos, con un valor de 84.10%. En este umbral destaca el *VPN* bajo, en ambos modelos. Esto indica que, cuando el modelo estime un sujeto como negativo (AHI menor a 1 e/h), aproximadamente en la mitad de los casos el sujeto sí presentará la enfermedad (AHI mayor a 1 e/h).

A pesar de que la diferencia de las métricas de los dos modelos no sea pronunciada, el modelo ResNet50 tiene un rendimiento en general ligeramente superior al de ResNet50 + Retina. Esto se refleja en un *Acc* 4-clases de 68.30% frente 65.36% y un coeficiente kappa de 0.5326 frente 0.4887. Este comportamiento sugiere que el pre-entrenamiento con imágenes de retina, a pesar de compartir similitudes cromáticas con los

espectrogramas, no aporta una ventaja en el aprendizaje de las características relevantes para esta tarea específica.

5.2 ESTIMACIÓN DEL AHI

Para poder discutir de forma completa el rendimiento de los modelos de este trabajo también se debe atender a la forma en la que estima el AHI, debido a que éste se usa para diagnosticar y estadificar la apnea.

Atendiendo al ICC de la Tabla 6 y la Tabla 8, ambos modelos presentan un ICC elevado, lo que indica una buena concordancia entre las predicciones y los valores reales. Para el caso de ResNet50, se obtiene un valor de 0.891, que, aunque no llegue a ser mayor que 0.9, se aproxima bastante. En el caso de ResNet50 + Retina se alcanza un ICC superior que el anterior de 0.911. Esto quiere decir que, aunque el modelo ResNet50 sea mejor que el ResNet50 + Retina en cuanto a métricas y detección de verdaderos positivos y negativos, esta segunda es mejor en acuerdo global y fiabilidad.

En cuanto a los diagramas de *Bland-Altman* de las Figuras 9 y Figura 11, se puede ver que ambos modelos tienden a subestimar el AHI en comparación con el AHI real. Esto se comprueba atendiendo al sesgo negativo de ambos. Para ResNet50, el sesgo es de -0.85 y para ResNet50 + Retina, de -0.58. Esta diferencia indica una menor subestimación en el caso del segundo modelo. Observando los límites de acuerdo, para ResNet50, el superior es 5.29 y el inferior -7, mientras que para ResNet50 + Retina el superior es 5.24 y el inferior de -6.40. Esto representa que la diferencia que hay entre el AHI real y el estimado, en el 95% de los casos, cae entre estos valores. Ambos modelos tienen límites de acuerdo bastante similares, siendo algo menor para ResNet50 + Retina. Las diferencias son tanto positivas como negativas y son lo suficientemente grandes como para ser consideradas relevantes en la aplicación de los modelos.

Finalmente, se observa que la dispersión de los puntos aumenta a medida que lo hace el valor promedio del AHI. Esto sugiere que, para valores más elevados de AHI, los errores de estimación también son mayores. En particular, se detectan casos extremos de subestimación, especialmente en el modelo ResNet50, donde algunas diferencias alcanzan más de 30 unidades de AHI.

5.3 INTERPRETACIÓN MEDIANTE SHAP

Una vez generadas las explicaciones SHAP correspondientes a las mejores y peores estimaciones del AHI para cada clase de severidad, se procede a su análisis interpretativo. Para ello, cabe recordar que los mapas de calor generados asocian el color rojo a valores SHAP positivos, blancos al cero y azules a valores SHAP negativos. Además, se emplea un mapa adicional que conserva el color rojo a positivos, gris al cero y azules a negativos, pero sobre el espectrograma original en escala de grises. Este segundo mapa sirve para aplicarlo sobre las imágenes y hacer zoom sobre regiones para una mejor interpretación.

Debido a que las señales de SpO_2 presentan una variación lenta y han sido remuestreadas a 1 Hz, los espectrogramas correspondientes se concentran en el rango de frecuencias entre 0 y 0.5 Hz.

5.3.1 CLASE NO-AOS

Para la clase No-OSA (AHI ≤ 1), la mejor estimación para un AHI real de 0.533 e/h es un AHI de 0.538 e/h. En la Figura 11 se observa que la señal en la mayoría del tiempo toma valores cercanos al cero o negativos. Esto significa que esas zonas de la señal no están contribuyendo al aumento del AHI al tratarse de un AHI real bajo. Si bien existen tres marcas de apnea e hipopnea, estas no son detectadas con intensidad significativa. Estas zonas se han podido confundir como contribuyentes debido a fluctuaciones de la señal. Sobre los colores leves rojizos se puede ver que la señal toma una forma más rizada que en el resto del tiempo, que es más estable y plana. En la Figura 19 derivada de la Figura 11, se puede encontrar una zona con color rojo intenso en el recuadro 'a' , donde el modelo está tomando la señal como información relevante para el cálculo del AHI. Sin embargo, no corresponde a una marca de apnea o hipopnea, posiblemente se deba a una desaturación no asociada a un evento. Por el contrario, el recuadro 'b' destacado en esta figura se trata de un ejemplo de evento de apnea que no es detectado con tanta intensidad, aún pudiendo distinguir una leve tonalidad roja.

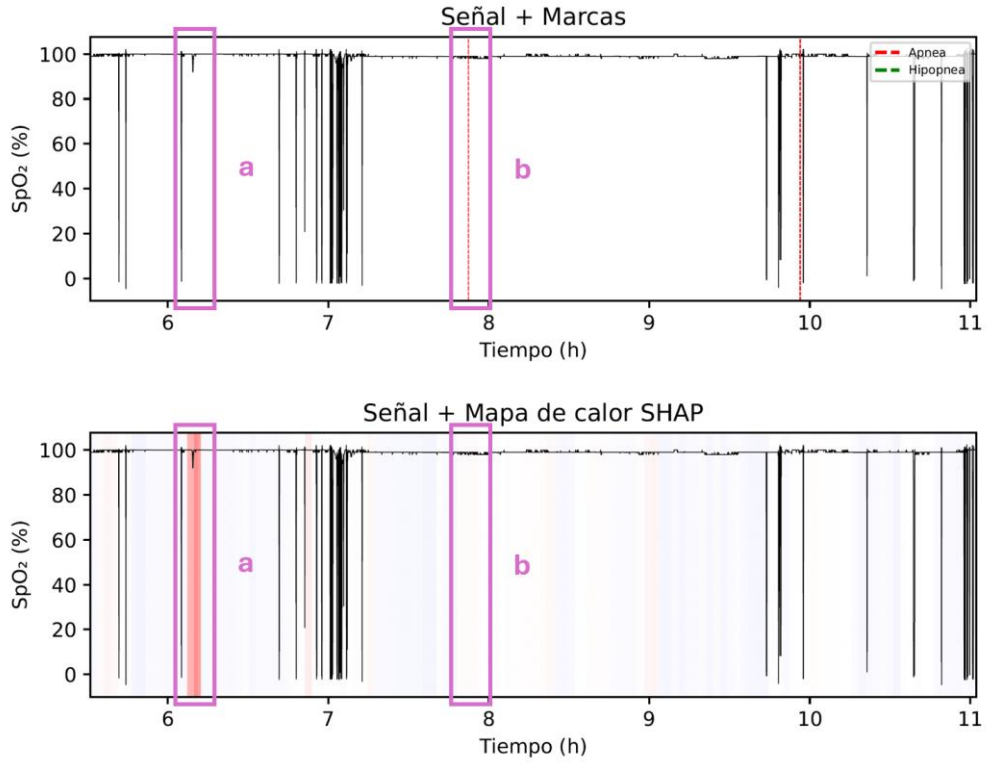

Figura 19. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase No-AOS con la mejor estimación del AHI.

Al observar el mapa de calor generado para el espectrograma en la Figura 11, podemos ver cómo existen zonas leves rojas que se encuentran solo en las bajas frecuencias. En la Figura 20 se hace zoom sobre el espectrograma con el mapa. En esta nueva imagen podemos comprobar el suceso mencionado anteriormente en recuadros verdes destacados. El recuadro 'c' corresponde a la marca de apnea del recuadro 'b' de la Figura 19. Se observa cómo existen valores SHAP positivos que se distribuyen solo en las frecuencias <0.025 Hz. En el caso del recuadro 'c' puede deberse al aumento de potencia en las frecuencias bajas ligado a descensos de la SpO_2 lentos y prolongados provocados por la AOS. En el resto de los recuadros, como se trata de zonas libres de marcas, puede deberse a patrones de fluctuación lentos de la propia señal que el modelo considere importantes.

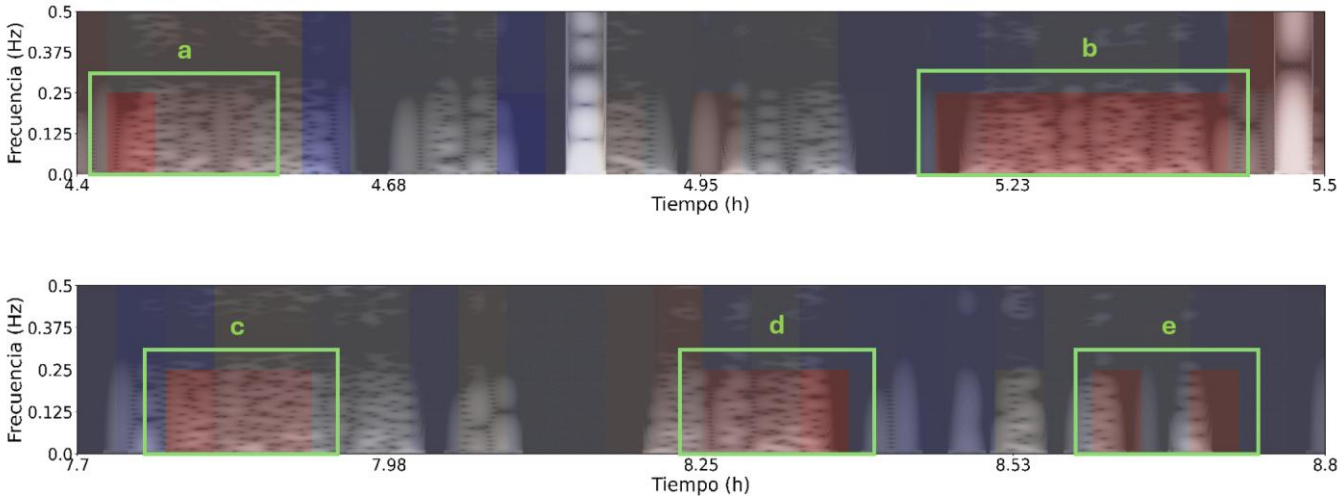

Figura 20. Zoom realizado en una zona del espectrograma de la mejor estimación para No-AOS con mapa SHAP por encima.

Pasando a la peor estimación, nos encontramos con un AHI de 2.802 e/h frente un AHI real de 0.606 e/h. En este caso, como podemos observar en la Figura 12, no hay tantas tonalidades azules que contribuyan con valor SHAP negativo, sino que la mayoría de las zonas están cerca de valores SHAP próximos a cero. Esto, junto a la detección incorrecta de algún evento, puede aumentar erróneamente el AHI.

En el recuadro 'a' de la Figura 21 y los recuadros 'c' y 'd' de la Figura 22, derivadas ambas de la Figura 12, podemos comprobar cómo la red vuelve a detectar como relevantes algunas caídas en la oxigenación no asociadas a eventos apneicos. Sin embargo, en el recuadro 'b' de la Figura 21, detecta de forma correcta una marca de apnea, aunque sin mucha intensidad.

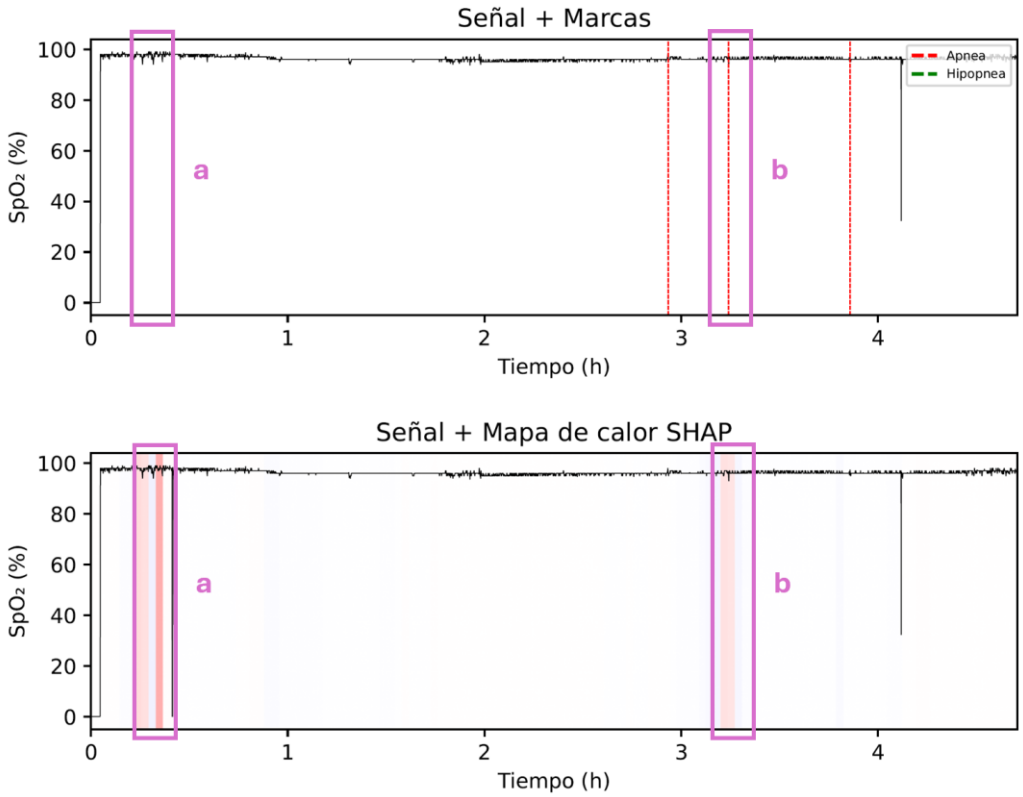

Figura 21.Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase No-AOS con la peor estimación del AHI.

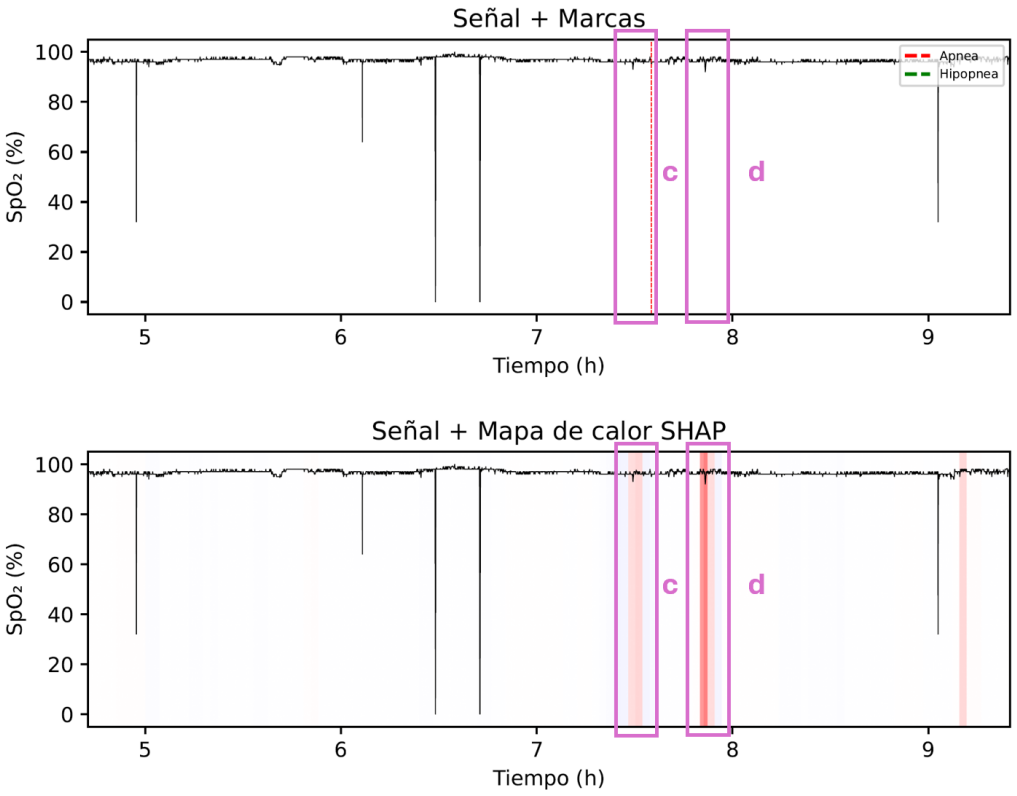

Figura 22.Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase No-AOS con la peor estimación del AHI.

En la Figura 23, en el recuadro ‘a’, se contempla en el espectrograma un ejemplo de confusión con una posible desaturación no asociada a la AOS correspondiente al recuadro ‘b’ de Figura 22. Por otro lado, en la imagen inferior, se encuentra la detección de una marca real de apnea correspondiente al recuadro ‘b’ de la Figura 21. Si bien la presencia de importancia en la banda inferior (<0.025 Hz) es esperable debido a la naturaleza lenta de las apneas, también se registra potencia en frecuencias superiores, lo cual podría asociarse a un aumento de potencia debido a esfuerzos respiratorios o cambios dinámicos de las señales derivados de los eventos. Además, también se contempla cómo las zonas rojas más intensas están rodeadas de zonas azules. Esto tiene sentido, pues se detecta como valor positivo el posible evento y como negativo sus alrededores, delimitándolo temporalmente.

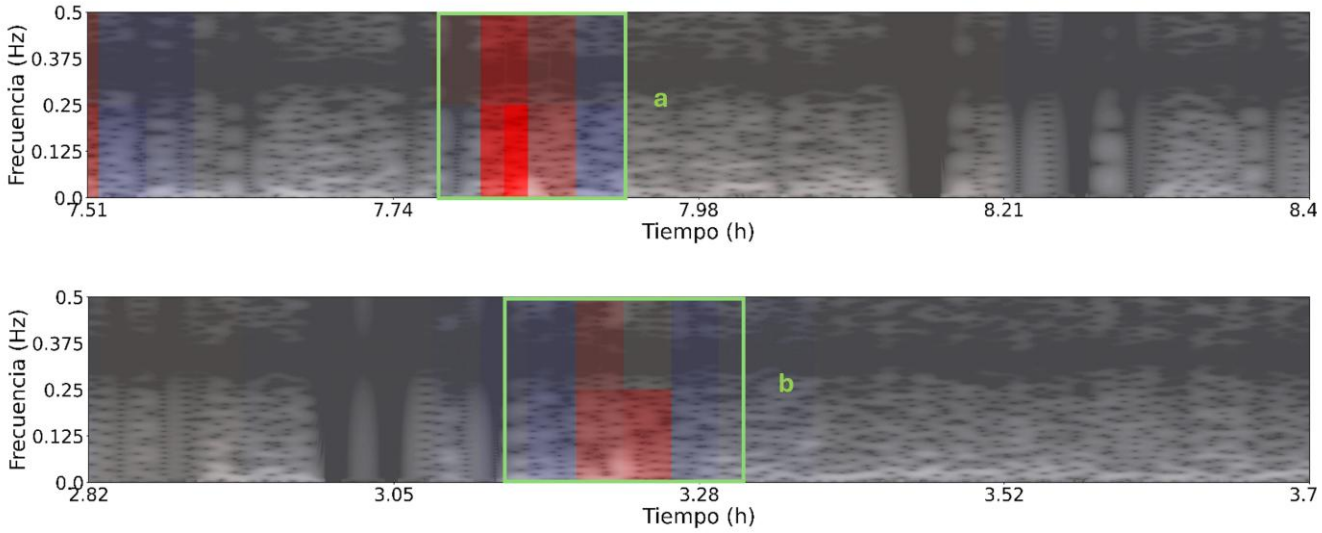

Figura 23. Zoom realizado en una zona del espectrograma de la peor estimación para No-AOS con mapa SHAP por encima.

5.3.2 CLASE AOS LEVE

En la clase AOS leve ($1 \leq \text{AHI} < 5$), la mejor estimación de AHI es 1.347 e/h siendo AHI real 1.328 e/h. En la Figura 13 se observa que algunas marcas de apneas e hipopneas se detectan de forma correcta frente a otras que pasan desapercibidas. Además, se vuelve a dar importancia a algunas caídas en la oxigenación no asociadas a eventos.

En el recuadro ‘a’ de la Figura 24 y el recuadro ‘d’ de la Figura 25, derivadas ambas de la Figura 13, se observa como aumenta el AHI debido a picos de bajadas en la saturación. Los recuadros ‘b’ y ‘c’ de la Figura 24 corresponden a unas marcas que sí se han asociado a valor positivo correctamente.

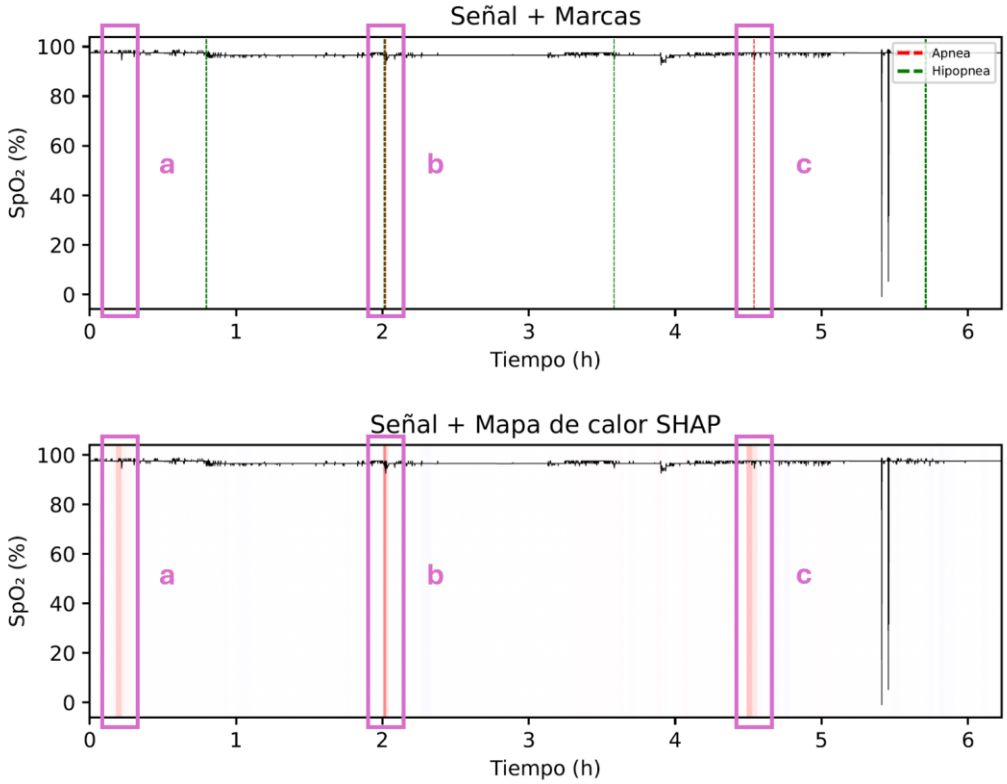

Figura 24. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS leve con la mejor estimación del AHI.

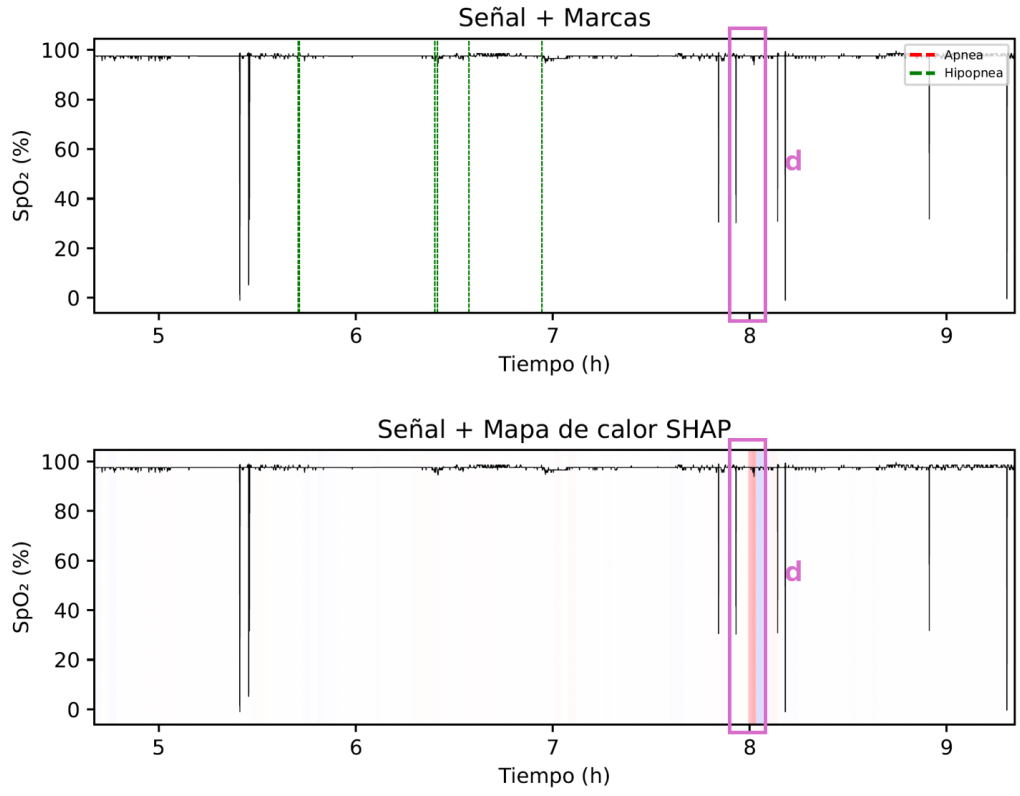

Figura 25. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS leve con la mejor estimación del AHI.

En la Figura 26, correspondiente al espectrograma ampliado, vemos una marca de hipopnea seguida de una marca de apnea breve correspondiente al recuadro ‘b’ de la Figura 24. Como se ha comentado, las apneas se reflejan en una variación lenta de la señal de SpO_2, donde se disminuye progresivamente la saturación.

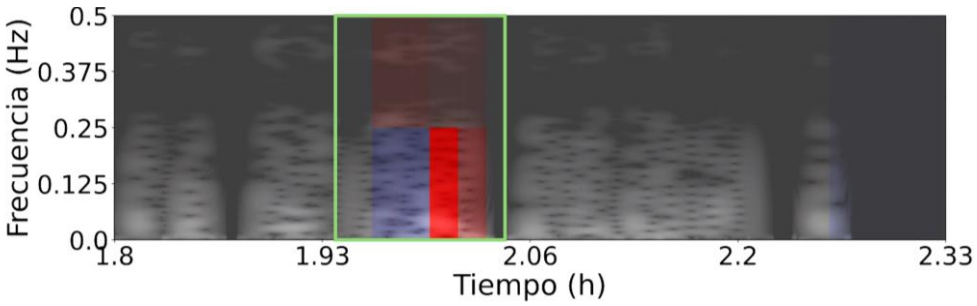

Figura 26. Zoom realizado en una zona del espectrograma de la mejor estimación para AOS leve con mapa SHAP por encima.

En la peor estimación de esta clase el AHI estimado es de 11.955 e/h, mientras que el real es de 4.855 e/h. En la Figura 15 se identifican múltiples zonas rojizas asociadas a regiones sin eventos clínicos marcados, lo que sugiere errores de interpretación.

En la Figura 27, derivada de la Figura 14, en el recuadro ‘a’ vemos una serie de marcas en algunos casos bien diferenciadas. En el recuadro ‘b’ vemos marcas de apnea, pero con una intensidad demasiado leve. Cabe destacar que, por lo general, el modelo es capaz de aprender en todas las señales que los artefactos no son relevantes y no tienen relación con el AHI.

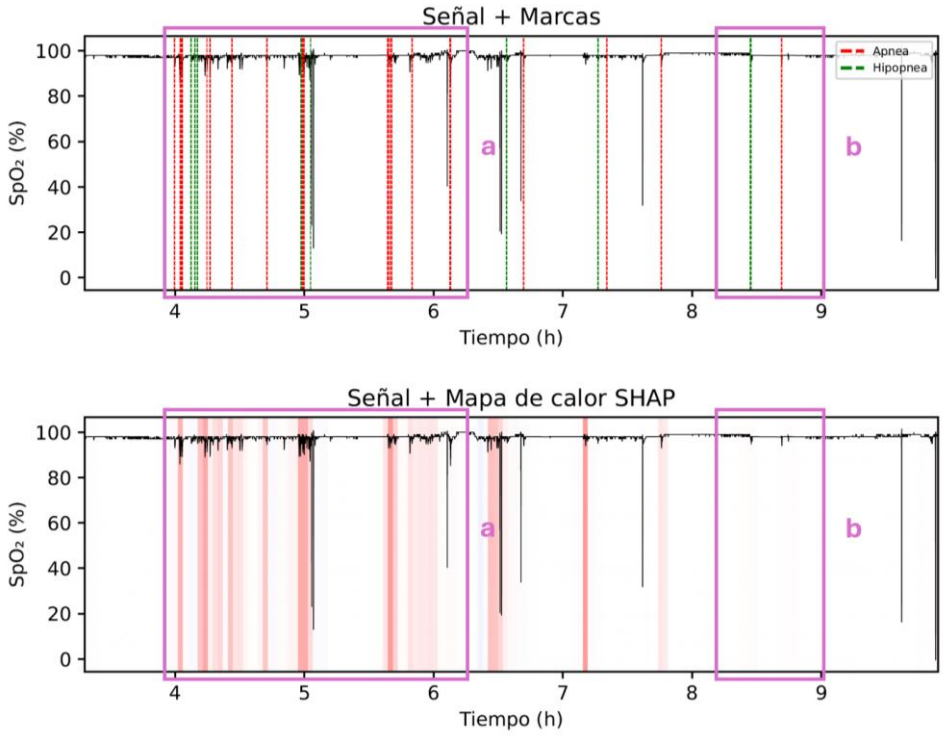

Figura 27. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS leve con la peor estimación del AHI.

En la Figura 28 vemos una parte espectral del recuadro ‘a’ de la Figura 27. Comprobamos que hay demasiadas zonas con color rojo que están aumentando erróneamente el AHI. Las zonas recuadradas en verde corresponden a las zonas donde ocurren temporalmente las marcas, las cuales contribuyen en la predicción. Sin embargo, las zonas intermedias en las que hay ausencia de apneas e hipopneas, también están contribuyendo con intensidad aumentando el error.

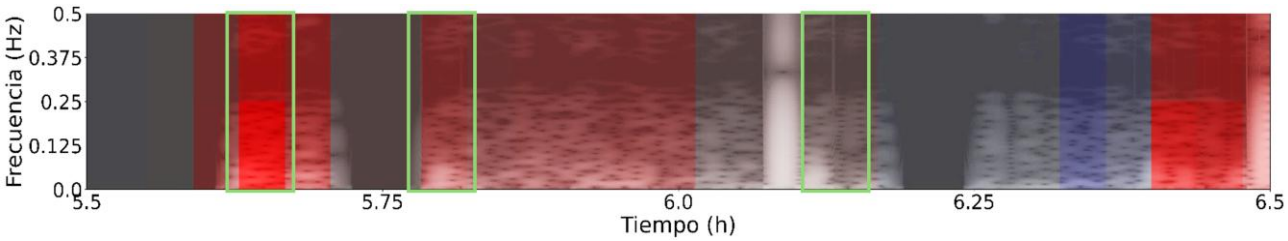

Figura 28. Zoom realizado en una zona del espectrograma de la peor estimación para AOS leve con mapa SHAP por encima.

5.3.3 CLASE AOS MODERADA

En la clase AOS moderada (5 ≤ AHI < 10), la mejor estimación (AHI real: 8.182 e/h; estimado: 8.050 e/h) refleja una alta precisión. La Figura 15 muestra múltiples marcas correctamente reconocidas por el modelo.

En la Figura 29, derivada de la Figura 15, destaca en el recuadro ‘a’ la detección de una serie de apneas e hipopneas agrupadas que ocurren de forma muy consecutiva.

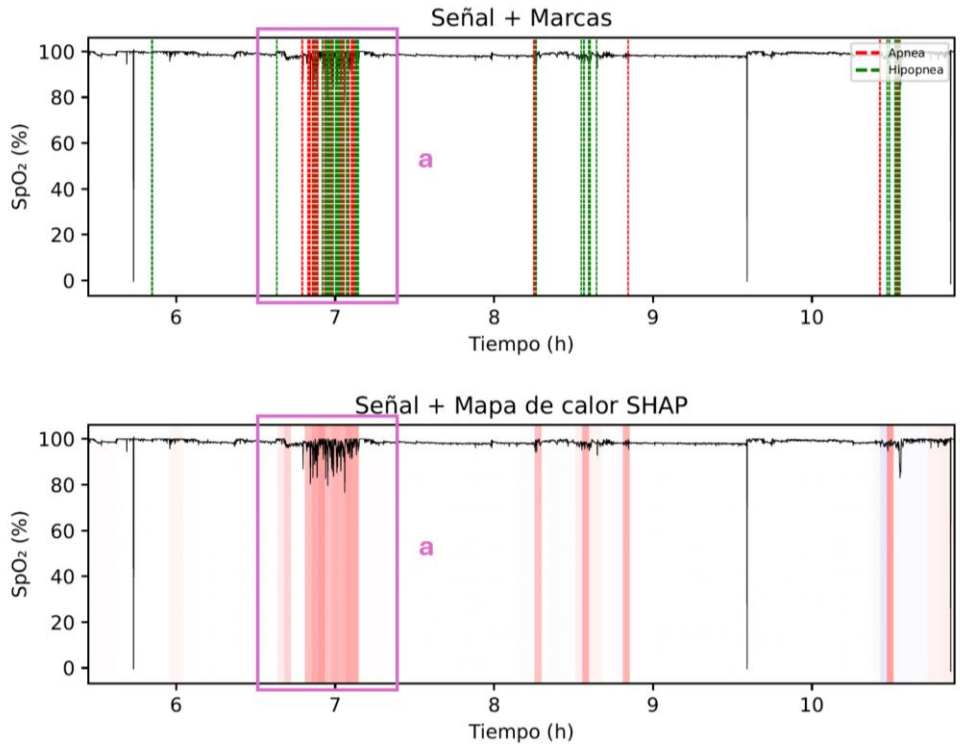

Figura 29. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS moderada con la mejor estimación del AHI.

En la Figura 30 podemos ver espectralmente el recuadro ‘a’ de la Figura 27. Esta imagen nos ayuda a confirmar lo siguiente. Las regiones del espectrograma que la red toma como importantes en la tarea de regresión tienen un aspecto en común. Estas regiones cuentan con líneas verticales de aumento de potencia que se extienden a lo largo de todo el eje de frecuencia. En algunas zonas se asocian valores SHAP positivos más grandes en bajas frecuencias (0-0.25 Hz), posiblemente debido a un aumento de potencia llamativo en estas zonas. Esto nos explica lo esperado, el aumento de potencia en frecuencias bajas es debido a los eventos apneicos de variación lenta. Aun así, la importancia en frecuencias altas se mantiene intensa debido posiblemente a lo comentado con anterioridad: el aumento de potencia también en altas frecuencias debido a esfuerzos respiratorios u otras variaciones dinámicas derivadas de los eventos. En la Figura 31, se muestra el fenómeno comentado para el espectrograma en en color.

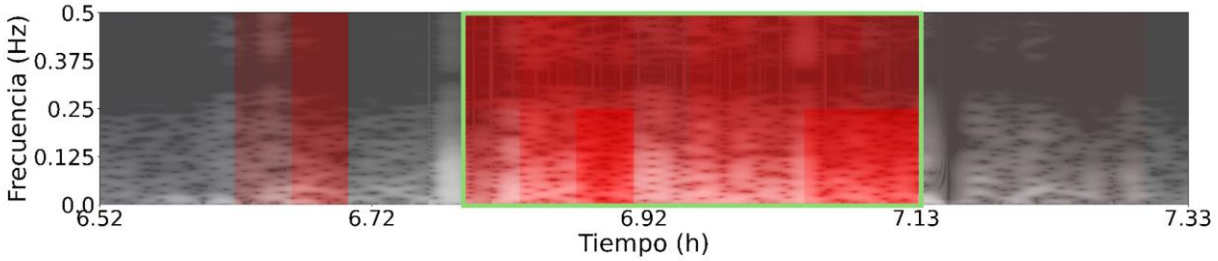

Figura 30. Zoom realizado en una zona del espectrograma de la mejor estimación para AOS moderada con mapa SHAP por encima

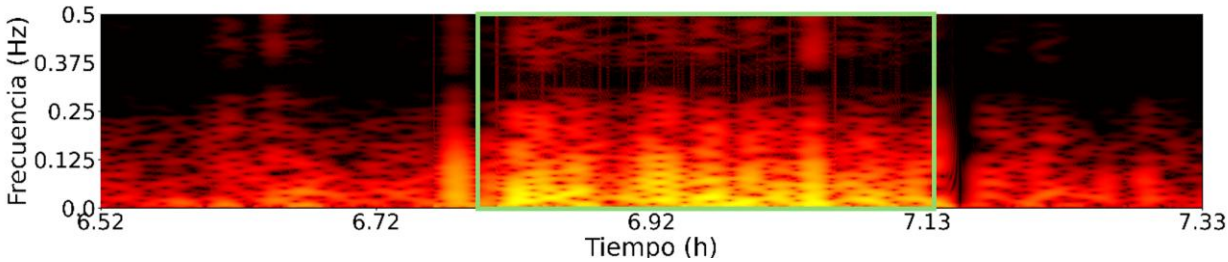

Figura 31. Zoom realizado en una zona del espectrograma en color de la mejor estimación para AOS moderada.

En el caso de la peor estimación para AOS moderada, se predice un AHI de 2.445 e/h y el real es 8.213 e/h. En la Figura 16 se aprecia una señal con múltiples marcas de hipopnea formando grupos y que, sin embargo, en la explicación, la mayoría son valores cercanos al cero con escasas zonas positivas.

La Figura 32, derivada de la figura 16, ilustra en el recuadro ‘a’ uno de los grupos de hipopneas que se asocia erróneamente a valores cercanos a 0, haciendo que el AHI se subestime bastante respecto al original.

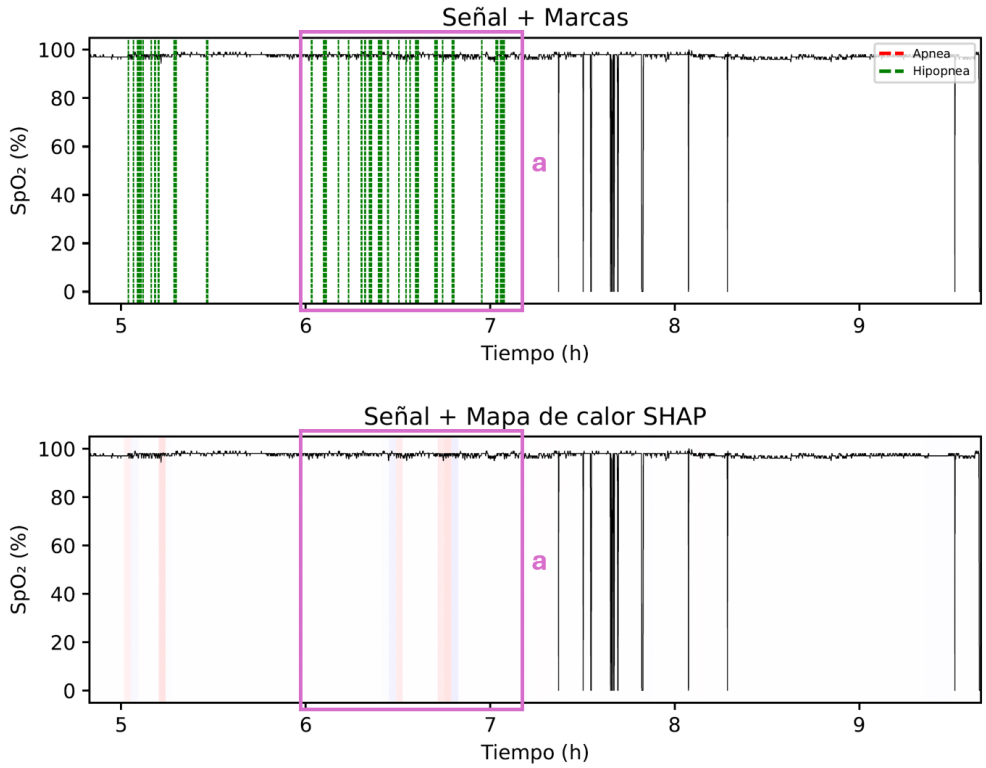

Figura 32. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS moderada con la peor estimación del AHI.

En la Figura 33 se hace zoom sobre el espectrograma en la zona correspondiente al recuadro 'a' destacado de la Figura 32. En este caso, la detección escasa de hipopneas puede deberse a la ausencia del fenómeno comentado en el anterior espectrograma. No se pueden distinguir con facilidad las bandas temporales de aumento de potencia que se extienden en todas las frecuencias.

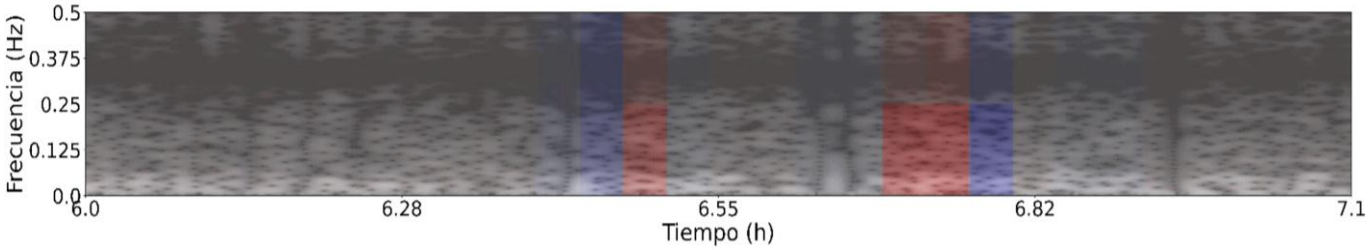

Figura 33. Zoom realizado en una zona del espectrograma de la peor estimación para AOS moderada con mapa SHAP por encima

5.3.4 CLASE AOS SEVERA

Por último, en la clase AOS severa (AHI ≥ 10) la predicción obtiene un AHI de 16.347 e/h frente al real de 16.396 e/h. En este caso, podemos ver que en la Figura 17 varios grupos de apneas e hipopneas se detectan como contribuyentes para el AHI, lo cual se refleja en una estimación buena.

En la Figura 34, derivada de la Figura 17, se destacan en los recuadros ‘a’, ‘b’ y ‘c’ algunos de estos grupos de marcas que el análisis SHAP asocia con valores SHAP positivos.

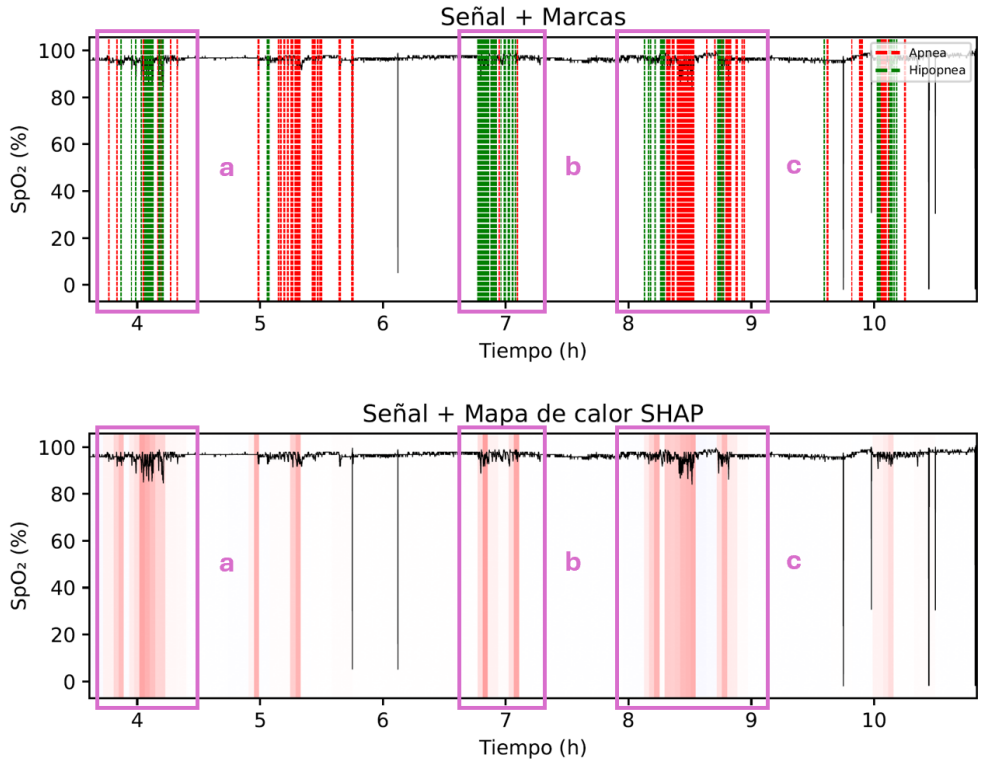

Figura 34. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS severa con la mejor estimación del AHI.

En la Figura 35, al aplicar zoom en el espectrograma en la zona correspondiente al recuadro ‘a’ de la figura 34, se observa que cuanto mayor es el número de bandas verticales transitorias de aumento de potencia en todas las frecuencias, más valor SHAP tiene la zona. En la Figura 36, se muestra este fenómeno para el espectrograma en color.

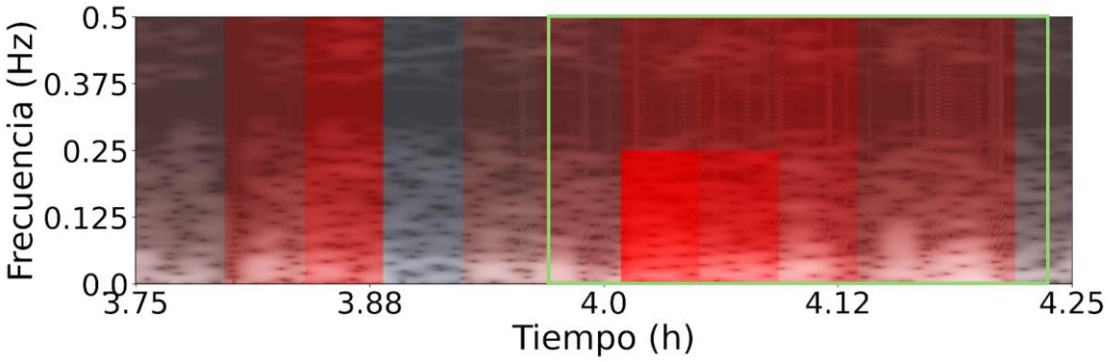

Figura 35. Zoom realizado en una zona del espectrograma de la mejor estimación para AOS severa con mapa SHAP por encima.

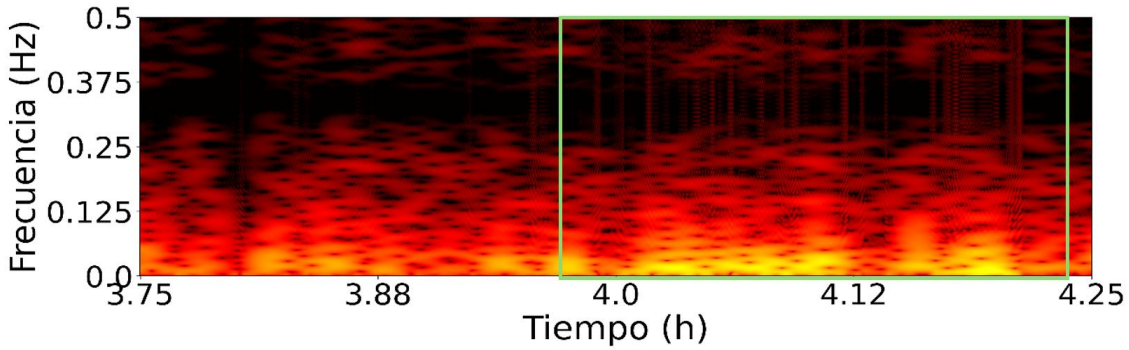

Figura 36. Zoom realizado en una zona del espectrograma en color de la mejor estimación para AOS severa.

En la peor estimación de esta clase contamos con una estimación AHI de 30.741 e/h y un AHI real de 62.313 e/h. En la Figura 18 se contempla como predominan los valores positivos debido a que se trata de un caso severo con un indice muy alto. Aun así, es la predicción con el mayor error debido a que muchos grupos de marcas consecutivas son ignorados.

En los recuadros 'a' y 'b' de la Figura 37 y los recuadros 'c' y 'd' de la Figura 38, derivadas de la Figura 18, podemos deducir un fenómeno adicional. En presencia de artefactos, las marcas que los rodean no son detectadas con tanta facilidad. Este comportamiento puede deberse a que la red ha aprendido que los segmentos de la señal afectados por artefactos no están relacionados con el AHI.

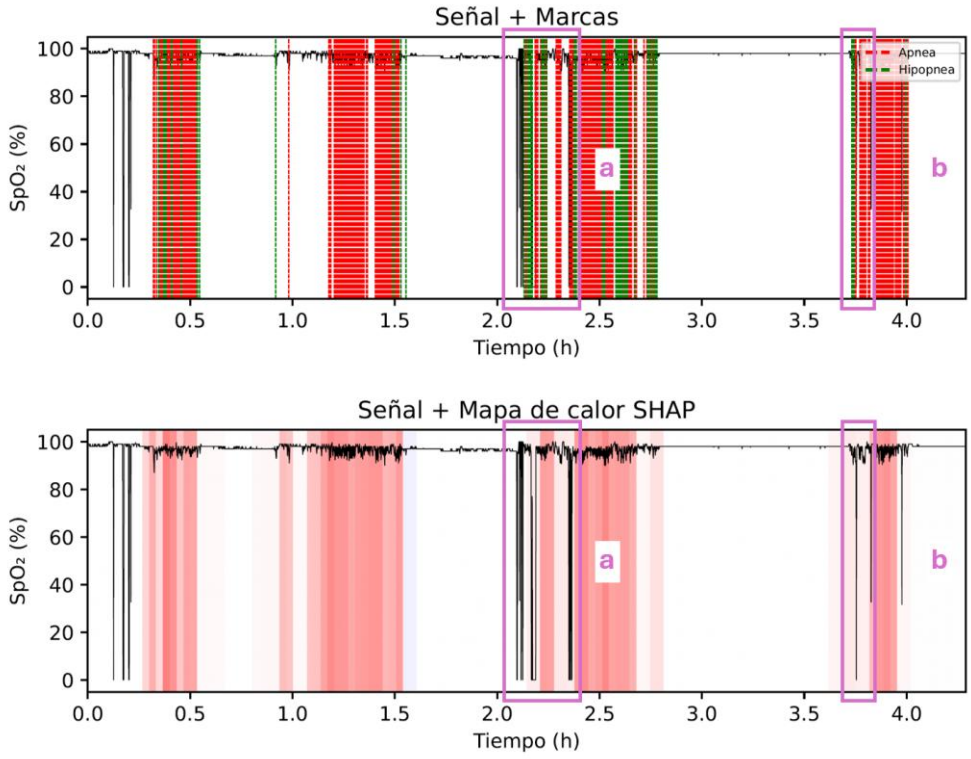

Figura 37. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS severa con la peor estimación del AHI.

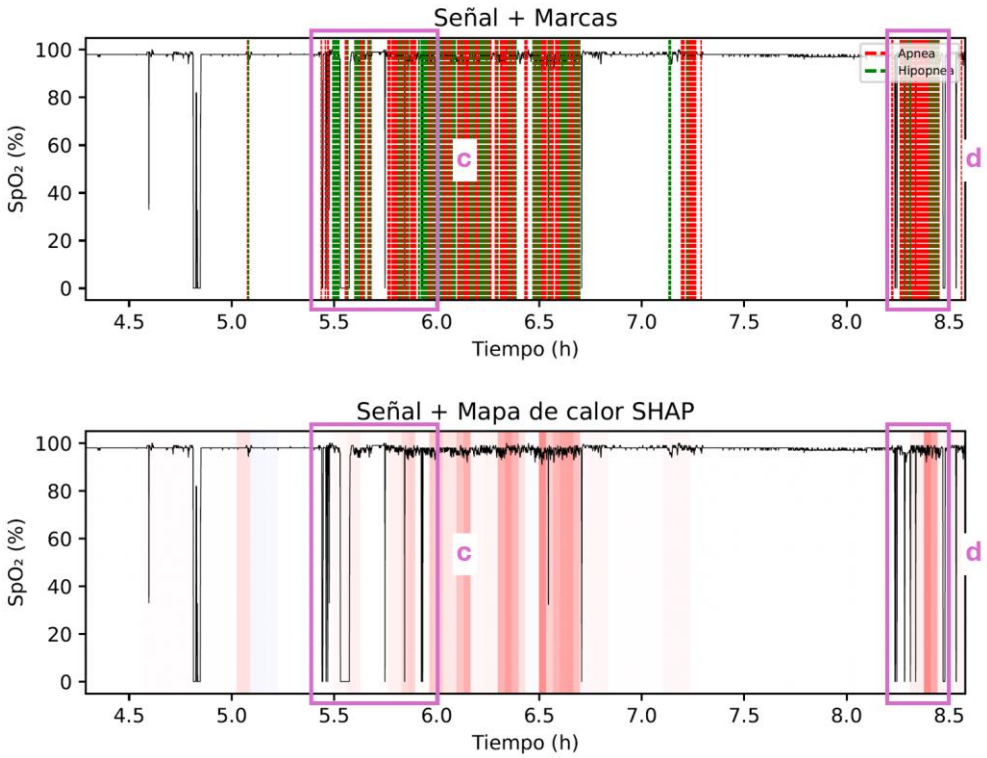

Figura 38. Ejemplos de sucesos destacados en el análisis SHAP para el sujeto de clase AOS severa con la peor estimación del AHI.

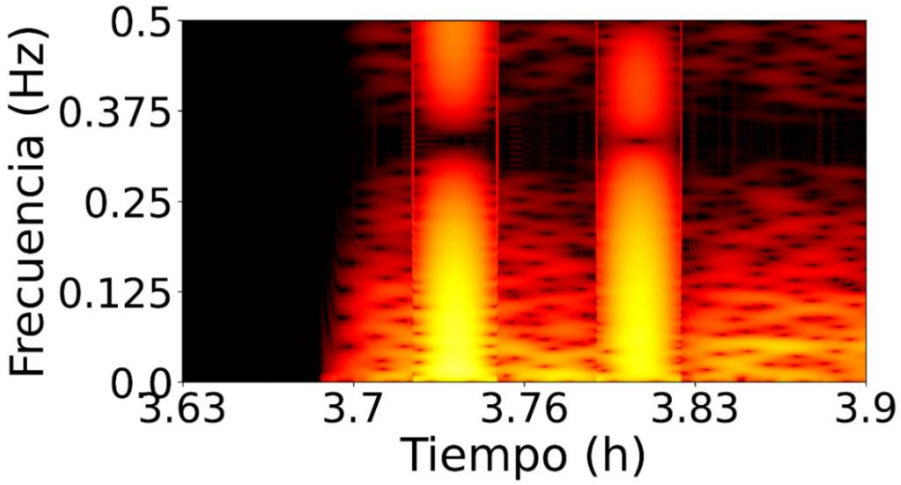

Figura 39. Zoom realizado en una zona del espectrograma en color de la peor estimación para AOS.

Al comparar todas las clases, se comprueba que sus análisis SHAP cumplen con lo comentado: cuanto más grave es la AOS, más marcas de apneas se ignoran y, por lo tanto, más se subestima el AHI. Aunque el error vaya aumentando a medida que el AHI es más grande, para las mejores estimaciones, el error se considera pequeño. Sin embargo, para las peores estimaciones, que comienzan en 2.802 para No-AOS y suben hasta 31.572 en AOS, se debe considerar bastante importante.

Esto puede guardar una relación con la elección de 2 minutos como duración de la ventana de la STFT. La ventana de 2 minutos, frente a duraciones inferiores, cuenta con una mayor resolución espectral pero una menor resolución temporal. Por esta razón, el modelo podría no reconocer eventos cortos y rápidos, especialmente en casos graves donde ocurren muchas apneas o hipopneas seguidas. Esto se vería reflejado en la subestimación del AHI aumentando el error para los casos más severos.

5.4 COMPARACIÓN CON OTROS ESTUDIOS

En esta sección se realiza una comparación entre el mejor modelo propuesto en este TFG, ResNet50 sin pre-entreno con imágenes de retina, con el resto de las investigaciones científicas existentes en la actualidad. Todas las investigaciones presentadas se centran en el diagnóstico automático de AOS a partir de diferentes señales biomédicas y se recogen en la Tabla 10.

Como se ha explicado anteriormente, la cantidad de estudios que usan imagen procedente de señal de SpO_2 para el diagnóstico de AOS infantil es escasa. La mayoría de los estudios son en población adulta y emplean otras señales entre las que destaca el ECG.

En cuanto a estos estudios de ECG en adultos, solo unos pocos emplean técnicas de XAI para mejorar la fiabilidad de las predicciones a partir de las explicaciones. Por ejemplo, en el estudio de Nasifoglu and Erogul, (2021), se predicen eventos de AOS en adultos en dos clases y se incorpora Grad-CAM. En los estudios de Akter et al., (2025) y Choudhury et al., (2025) también se trabaja con adultos, clasificando segmentos de 60 segundos en dos clases y aplicando Grad-CAM en el primero y LIME en el segundo. Estos estos estudios de Akter et al., (2025), Choudhury et al., (2025) y Nasifoglu and Erogul, (2021), se alcanzan precisiones de 98.79% 93.85% y 82.30% respectivamente. Estas exactitudes son superiores, pero difíciles de comparar con la señal y población empleada en este trabajo, debido a que el ECG es una señal que requiere más electrodos y tiene unas exigencias completamente diferentes.

En cuanto al estado del arte recogido para este trabajo, solo los estudios de *Mortazavi et al., (2023)* y Crowson et al., (2023), se realizan en población infantil. El trabajo de Mortazavi et al., (2023) utiliza señal de SpO_2 de la base de datos CHAT, para estimar el AHI y clasificar la AOS en cuatro clases. Para llevar a cabo la tarea, emplea espectrograma de segmentos de 20 minutos de la señal como imagen de entrada a la arquitectura InceptionV3, alcanzando un *Acc* del 74.11%. Cabe destacar que a diferencia de Mortazavi et al., (2023), en este trabajo se emplean los espectrogramas de la señal entera, lo que permite que la red aprenda de todas las características temporales y espectrales de una vez, y no pierda información sobre las relaciones entre estas. Por otra parte, en el estudio de Crowson et al., (2023) se utiliza la señal de presión de aire nasal para clasificar los eventos respiratorios en cuatro clases logrando un *Acc* del 70%. Los resultados de ambos estudios no alcanzan métricas tan altas en comparación con el resto de las señales aplicadas en población adulta.

Debido a que la señal, la técnica de transformación de la señal a imagen y los umbrales de AHI de Mortazavi et al., (2023) coinciden con los aplicados en este trabajo, podemos llevar a cabo una comparación de resultados más precisa. Para el umbral AHI=1 e/h, el modelo InceptionV3 de Mortazavi et al., (2023) presenta precisión y sensibilidad superiores (*Acc* 85.8% frente 81.37% y *Se* 96.8% frente 84.10%). Sin embargo, la especificidad de 52.3% es significativamente más baja comparado con las que se obtiene en este trabajo con ResNet50, un 71.64%, mostrando un mayor equilibrio entre sensibilidad y especificidad. Con el umbral AHI=5 e/h la precisión de 90.50% del estudio de Mortazavi et al., (2023) es muy similar a la conseguida con el modelo ResNet50 propuesto, de 90.52%. En las métricas de especificidad y *VPP,* el modelo InceptionV3 alcanza valores ligeramente superiores (*Sp* 96.60% frente 95.81% y *VPP* 90.40% frente 88.41%). En las restantes, sensibilidad y *VPN*, ocurre al contrario y ResNet50 propuesto tiene mejor desempeño (*Se* 76.00% frente 78.02% y *VPN* 90.60% frente 91.15%). Para el umbral AHI=10 e/h, al igual que en este trabajo, Mortazavi et al., (2023), alcanza la máxima precisión. Para este umbral de AHI, el modelo ResNet50 propuesto consigue una mejor sensibilidad (78.57% frente 70%). En cuanto al *Acc* (95.10% frente 96.4%) y la *Sp* (97.73% frente 100%), el presente trabajo, no supera al estudio de Mortazavi et al., (2023) pero estas métricas son igualmente altas. Además de los resultados, en esta comparación destacan algunas diferencias en la metodología. En primer lugar, Mortazavi et al., (2023) emplea fragmentos de la señal, mientras que en este TFG se emplea la señal completa, evitando la segmentación artificial de la señal y la pérdida de relaciones entre componentes temporales. En segundo lugar, el grupo de test del presente trabajo contiene 306 sujetos frente al grupo de test de Mortazavi et al., (2023) con 85 sujetos, lo que aporta mayor representación en la evaluación del modelo.

La diferencia más importante de este TFG, con respecto al estudio mencionado anteriormente y, por lo general, con el resto de los estudios, es aportar resultados con XAI. Esto permite entender las decisiones del modelo en función de las regiones de las imágenes y, sobre todo, conocer el motivo de los aciertos y fallos de la red.

Además de estas investigaciones, se añade la revisión sistemática del diagnóstico automático de AOS infantil con *machine learning* de (Gutiérrez-Tobal *et al.*, 2022). Este metaanálisis incluye 19 estudios con 4.767 niños y calcula métricas de rendimiento para los tres umbrales AHI más habituales, obteniendo los siguientes valores promedio: AHI=1 e/h: *Se*=92.10%, *Sp*=38.60%, AHI=5 e/h: *Se*=76.20%, *Sp*=85.10%. y AHÍ=10 e/h: *Se*=68.20%, *Sp*=95.80%.

Comparando con los valores globales presentados en Gutiérrez-Tobal et al., (2022), obtenidos del estado del arte incluido en el meta-analisis, el modelo ResNet50 de este TFG presenta un rendimiento competitivo. Para el umbral AHI=1e/h, el modelo ResNet50 presenta una sensibilidad cercana al promedio (81.37% frente 92.10%) y una especificidad superior (71.64% frente 38.60%). Para el umbral AHI=5 e/h, los resultados de este TFG superan los promedios en sensibilidad (78.02% frente 76.20%) y en especificidad (95.81% frente 85.10%). Para finalizar, en el umbral AHI=10 e/h, de nuevo

se mantienen valores superiores en los resultados de ResNet50 tanto en sensibilidad (78.57% frente 68.20%) como en especificidad (97.73% frente 95.80%).

Es interesante incorporar a la comparación los estudios de Vaquerizo-Villar et al., (2021), incluido en el meta-análisis de Gutiérrez-Tobal et al., (2022), así como el estudio de Jiménez-García et al., (2022). Ambos estudios se centran en el análisis de la señal de SpO_2 (en el caso de Jiménez-García et al., (2022) combinada con la señal de flujo aéreo) para la estimación del AHI y la clasificación de la AOS en población pediátrica. A diferencia del modelo ResNet50 propuesto, que emplea imágenes generadas a partir de la señal de SpO_2, estos dos estudios emplean directamente las señales como entrada a la red. No obstante, es especialmente interesante la comparación con los resultados de este TFG debido a que se centran en población infantil, emplean la misma base de datos (CHAT) y clasifican la severidad de la AOS con los mismos umbrales de AHI.

Comparando el modelo ResNet50 propuesto, con los resultados de la 2D CNN de Jiménez-García et al., (2022), para el umbral AHI=1 e/h, ambos obtienen la sensibilidad y el *VPP* más alto de todos los umbrales. De las sensibilidades de los dos modelos, ResNet50 alcanza un valor superior (84.10% frente 82.43%), minimizando en mayor medida los falsos negativos. Sin embargo, la especificidad del modelo ResNet50 es considerablemente inferior a la obtenida con la 2D CNN (71.64% frente 92.54%), lo que implica que el modelo de Jiménez-García et al., (2022) es más eficaz en la detección de pacientes sanos. En el resto de las métricas, el modelo ResNet50 no logra alcanzar los resultados de la 2D CNN: *Acc* 81.37% frente 82.43%, *VPP* 91.36% frente 97.52%, *VPN* 55.8% frente 59.62%, *LR*+ 0.22 frente 0.19, *LR*- 0.22 frente 0.19.

Para el umbral AHI=5 e/h todas las métricas de Jiménez-García et al., (2022) mantienen un carácter superior comparadas con las obtenidas en este trabajo: *Acc* (93.46% frente 90.52%), *Se* (80.22% frente 78.02%), *Sp* (99.07% frente 95.81%), *VPN* (92.21% frente 91.15%), y *LR*- (0.20 frente 0.23). Además, destaca particularmente la diferencia en el *VPP* (97.33% frente 88.75%) y *LR*+ (86.24 frente 18.62) que implica una mayor certeza a la hora de confirmar los diagnósticos positivos. Aun así, los resultados del modelo ResNet50 son competitivos al emplear únicamente la señal de SpO_2.

Por último, para el umbral AHI=10 e/h, ambos estudios obtienen sus mejores resultados de *Acc* y *VPN*. El modelo propuesto ResNet50 presenta métricas ligeramente superiores que la 2D CNN en precisión (95.10% frente 94.44%), sensibilidad (78.57% frente 71.43%), *VPN* (96.63% frente 95.57%) y *LR*- (0.22 frente 0.29). En las tres métricas restantes, el estudio de Jiménez-García et al., (2022) presenta valores algo superiores a ResNet50: *Sp* 98.11% frente 97.73%, *VPP* 85.71% frente 84.62% y *LR*+ 37.71 frente 34.61. Esto sugiere que, en los casos más severos, ambos modelos muestran un rendimiento elevado, con diferencias pequeñas entre uno y otro.

En cuanto a la comparación de los resultados con la CNN del estudio de Vaquerizo-Villar et al., (2021), para el umbral AHI=1 e/h, el modelo ResNet50 propuesto obtiene mejores resultados de precisión (81.37% frente 77.06%), sensibilidad (84.10% frente 71.20), *VPP* (91.36% frente 72.40) y *LR*- (0.22 frente 0.35). En las métricas *Sp* (81.80% frente

71.64%) *VPN* (81% frente 55.81%) y *LR*+ (3.92 frente 2.97) los valores son mejores para la CNN de Vaquerizo-Villar et al., (2021). Como ResNet50 es más sensible, tendrá mayor capacidad de detectar casos positivos de AOS, mientras que la CNN al ser más específica, clasificará mejor los sujetos sin patología.

Para el umbral AHI=5 e/h los resultados de la CNN son mejores que los del modelo ResNet50 propuesto: *Acc* (97.40% frente 90.52%), *Se* (83.80% frente 78.02%), *Sp* (100% frente 95.81%), *VPP* (100% frente 88.75%), *VPN* (97.00% frente 91.15%) y *LR*- (0.16 frente 0.23). EL valor LR+ no es comparable puesto que no se conoce en el estudio de Vaquerizo-Villar et al., (2021). La comparación sugiere que, a medida que aumenta el umbral de AHI, el modelo de Vaquerizo-Villar et al., (2021) obtiene un desempeño mejor, destacando de entre todos los resultados, la especifidad de 100%, logrando la perfección en la detección de casos negativos.

Para el umbral AHI=10 e/h, tanto la red CNN como la ResNet50, obtienen los mejores valores de precisión, *VPN*, *LR*+ y *LR*-. Además, al igual que para el umbral AHI=5 e/h, todos los resultados de Vaquerizo-Villar et al., (2021) se mantienen por encima del modelo propuesto en este trabajo: *Acc* (97.80% frente 95.10%), *Se* (83.90% frente 78.57%), *Sp* (99.30% frente 97.73%), *VPP* (92.90% frente 84.62%), *VPN* (98.20% frente 96.63%) y *LR*- (0.16 frente 0.22). Destaca la diferencia en *LR*+ (117.84 frente 37.71), lo que significa que, cuando la CNN del estudio Vaquerizo-Villar et al., (2021) detecta un caso positivo, la probabilidad que de que realmente sea positivo frente a uno negativo es mucho mayor que cuando lo predice el modelo ResNet50.

Aunque ResNet50 no alcance los valores de algunas métricas de los estudios de Vaquerizo-Villar et al., (2021) y Jiménez-García et al., (2022), ofrece un buen balance entre sensibilidad y especificidad, especialmente en los umbrales altos (5 e/h y 10 e/h), donde supera los valores promedio del meta-análisis. Además, un aspecto destacable es que, para el umbral AHI=1 e/h, el modelo ResNet50 propuesto en este trabajo, comparado con los dos anteriores, logra la sensibilidad más alta. Este resultado es relevante en un contexto clínico ya que a partir de 1 e/h, se considera AOS leve y una sensibilidad más alta implica menor número de falsos negativos. Por lo tanto, el modelo ResNet50 minimiza el riesgo de que pacientes que padecen AOS sean clasificados como sanos. Esto es decisivo tanto en cribados como en diagnóstico temprano, donde es importante detectar el mayor número de casos positivos de AOS, para un seguimiento y tratamiento precoz.

ESTUDIO	POBLACIÓN	SEÑAL	ARQUITECTURA	XAI	UMBRAL AHI (e/h)	Acc (%)	Se (%)	Sp (%)	VPP (%)	VPN (%)	LR+	LR-
Modelo ResNet50 propuesto	Infantil	SpO_2	ResNet50	SHAP	1	81.37	84.10	71.64	91.36	55.81	2.97	0.22
					5	90.52	78.02	95.81	88.75	91.15	18.62	0.23
					10	95.10	78.57	97.73	84.62	96.63	34.61	0.22
(Jiménez-García *et al.*, 2022)	Infantil	SpO_2 + flujo aéreo	2D CNN	-	1	84.64	82.43	92.54	97.52	59.62	11.05	0.19
					5	93.46	80.22	99.07	97.33	92.21	86.24	0.20
					10	94.44	71.43	98.11	85.71	95.57	37.71	0.29
(Vaquerizo-Villar *et al.*, 2021)	Infantil	SpO_2	CNN	-	1	77.60	71.20	81.80	72.40	81.00	3.92	0.35
					5	97.40	83.70	100	100	97.00	N.D	0.16
					10	97.80	83.90	99.30	92.90	98.20	117.84	0.16
(Mortazavi *et al.*, 2023)	Infantil	SpO_2	InceptionV3	-	1	85.80	96.80	52.30	86.10	84.60	-	-
					5	90.50	76.00	96.60	90.40	90.60		
					10	96.40	70.00	100	100	96.10		
(Crowson *et al.*, 2023)	Infantil	Presión nasal	CNN	-	-	70.00	-	-	-	-	-	-
(Akter *et al.*, 2025)	Adultos	ECG	CNN	Grad-CAM	-	98.79	98.93	98.93	-	-	-	-
(Choudhury *et al.*, 2025)	Adultos	ECG	CNN	LIME	-	93.85	93.42	94.30	-	-	-	-
(Nasifoglu and Erogul, 2021)	Adultos	ECG	CNN	Grad-CAM	-	82.30	83.22	82.27	82.95	-	-	-

Tabla 10. Comparación de los resultados del modelo ResNet50 con otros estudios.

5.5 LIMITACIONES

Durante el desarrollo de este trabajo se han identificado diferentes limitaciones que han podido influir en los resultados.

La primera limitación que se ha encontrado es la cantidad de datos con los que se ha trabajado. Aunque la base de datos CHAT tenga la suficiente información sobre la señal de SpO_2 de niños con AOS, un posible conjunto de información aún más extenso podría haber ayudado a aumentar la generalización del modelo.

Asimismo, siguiendo en la línea del conjunto de datos empleado, una representación más balanceada de cada tipo de apnea también podría haber resultado más positiva en el entrenamiento de la red. En este estudio, la clase AOS leve contaba aproximadamente con entre el doble y el triple de imágenes más que el resto de las clases. Aunque se hayan repetido durante el entrenamiento imágenes en las clases minoritarias, un mayor equilibrio u otras formas de aumento de datos podría haber supuesto un aumento en el rendimiento del modelo.

Otra limitación que se ha presentado en variedad de ocasiones es el tamaño de imagen. Debido a que los espectrogramas se generaron de toda la señal completa, la dimensión resultante fue demasiado elevada. No se trabajó con varios segmentos de cada imagen, sino que se empleó la imagen entera como entrada a la red. Aún redimensionando y unificando esta entrada a 120×16000 píxeles, se trata de un tamaño inusual para ResNet50, tanto entrenada con *ImageNet* como pre-entrenada con retina.

En lo referente a la técnica SHAP, también se identifican determinadas limitaciones. El tamaño de las imágenes es inusual para el análisis por lo que el cálculo de las explicaciones se ralentiza. Además, SHAP asume independencia entre cada parte de la señal, algo que al tratarse de una señal biomédica para el estudio del sueño no se cumple pues hay relaciones temporales. Asimismo, los mapas generados son difíciles de interpretar y no siempre se puede identificar por qué ciertas regiones reciben valores SHAP positivos o negativos. Por último, es importante destacar que, por problemas de capacidad de computación, solo se realizó el análisis SHAP con un grupo reducido de 8 espectrogramas, lo que limita la extracción de conclusiones.

CAPÍTULO 6: CONCLUSIONES Y LINEAS FUTURAS

6.1 CONTRIBUCIONES

Este TFG aporta una serie de contribuciones relevantes al ámbito del diagnóstico automático de la AOS en población pediátrica, entre las que destacan:

- A diferencia de la mayoría de los estudios de diagnóstico automático de la AOS a partir de imagen médica, realizados en población adulta, este trabajo se centra en niños. La prevalencia es importante en niños y los métodos desarrollados en adultos no se pueden aplicar al no ser adecuados en estos casos, ya que las características fisiológicas son diferentes. Por ello, en el estudio de la AOS, es necesario aportar información sobre esta población para lograr mejorar la calidad de vida de los pacientes y reducir tanto las secuelas como los gastos derivados de un diagnóstico tardío.
- En este estudio se emplean imágenes derivadas de la señal de SpO_2, proporcionando innovación y un nuevo campo de desarrollo. Por un lado, presenta la ventaja de usar los espectrogramas, aprovechando la información aprendida en redes pre-entrenadas con millones de imágenes. Por otro lado, esta innovación es más interesante al usar la señal de SpO_2. Esta señal se presenta como un método no invasivo y accesible, alternativo a la PSG y a señales biomédicas más estudiadas como el ECG.
- La aplicación de XAI en la mayoría de los estudios en esta área es escasa. La aplicación de SHAP no solo aporta novedad, sino que también brinda confianza al método diagnóstico. Las explicaciones de las predicciones aportadas ayudan a entender el modelo desarrollado, otorgando a nuestra propuesta de mayor transparencia y conocimiento de cara a futuras mejoras.
- Este trabajo aporta una comparación entre modelos pre-entrenados con diferentes tipos imágenes, lo que aporta información sobre la influencia de los colores y las características de las imágenes usadas en el rendimiento de las redes.
- Por último, llevar a cabo tanto una regresión como una clasificación subsiguiente a esa estimación, permite tener información tanto del índice AHI como del diagnóstico de la enfermedad en cuatro severidades distintas. Esto proporciona información más completa y personalizada de la enfermedad de cada paciente.

6.2 CONCLUSIONES

Los resultados obtenidos en este trabajo permiten concluir que el uso combinado de DL junto técnicas de XAI con imagen procedente de la señal SpO_2 son útiles en el diagnostico automático de AOS infantil. Gracias a la estimación del AHI y su posterior utilización en la clasificación de la enfermedad en cuatro severidades, es posible crear una herramienta innovadora y más accesible como alternativa al uso de la PSG tradicional.

El uso de *transfer learning* con ResNet50 demuestra un rendimiento superior a ResNet50 + Retina en la mayoría de las métricas. Ambos modelos muestran sus mejores resultados ante la AOS severa, con exactitudes elevadas. Por tanto, se concluye que el uso de pre-entrenamiento con imágenes de retina no repercute en un rendimiento superior, aunque los colores asignados al espectrograma fueran en rangos similares a los colores típicos empleados en imágenes de fondo de ojo

Por otro lado, encontramos que ResNet50 + Retina obtiene un ICC superior a ResNet50, lo que implica mayor acuerdo entre sus predicciones y los datos reales. Además, los diagramas *Bland-Altman* demuestran la tendencia de subestimar el AHI por parte de ambos modelos, siendo algo menor para ResNet50 + Retina. Esto determina que el modelo ResNet50 tiene mayor rendimiento en clasificación, pero ResNet50 + Retina estima el AHI con mayor fiabilidad. Aún así, los errores cometidos por ambos no se deben ignorar si se emplean en un contexto clínico.

La técnica de SHAP, ha permitido analizar en detalle las decisiones del mejor modelo de ResNet50 en la estimación del AHI, lo que aporta confianza en el diagnóstico. De forma general, el análisis SHAP asoció valores positivos a las zonas marcadas con eventos de apnea e hipopnea y negativos o cercanos al cero a zonas normales de la señal, lo que cumple con lo esperado.

A lo largo de las interpretaciones de los mapas de calor generados se concluye que algunas desaturaciones no asociadas a apneas o hipopneas son detectadas erróneamente como zonas contribuyentes en el AHI. Además, se asocian valores positivos a zonas donde la señal fluctúa frente a zonas estables de la señal.

En las mejores predicciones de AOS moderada y severa se comprobó con claridad un patrón en todos los espectrogramas. Las marcas son detectadas debido a un aumento de potencia en bandas temporales transitorias para las bajas frecuencias (0-0.25 Hz).

Los análisis SHAP y *Bland-Altman* muestran que el error en la subestimación aumenta a medida que aumenta el AHI, posiblemente debido a la complejidad y al elevado número de eventos consecutivos. Por lo tanto, la clase AOS severa, aún siendo la que obtiene mejores métricas, es la que más sufre la subestimación del AHI. No obstante, debido a que el error absoluto en las mejores estimaciones sigue siendo pequeño, se concluye que el modelo es eficaz para la detección de AOS.

6.3 LINEAS FUTURAS

Al llevar a cabo este trabajo, se ha comprobado que se abre la posibilidad a numerosas investigaciones nuevas. A continuación, se recogen los aspectos más importantes.

- Ampliación de la base de datos: Resultaría interesante el empleo de diferentes bases de datos que incluyan la señal SpO_2. Esto contribuiría a tener información más diversa al ser recogida con diferentes dispositivos, sobre diferentes poblaciones y bajo condiciones distintas. Asimismo, se podría probar la utilización de bases fusionadas para entrenar con un mayor conjunto de datos y mejorar el aprendizaje durante el entrenamiento.
- Optimización de hiperparámetros: En este trabajo solo se ha podido realizar una búsqueda exhaustiva de un único hiperparámetro, el número de neuronas de las capas *Dense*. Por esta razón, otra posibilidad interesante sería tratar de optimizar otros hiperparámetros como el *dropout*, el *learning rate* o el número de capas. Esto permitiría un ajuste más fino y posiblemente un aumento del rendimiento en la estimación del AHI.
- Segmentación de la señal de entrada: Las imágenes de entrada a la red tienen un tamaño de 120x16000 píxeles. Esta dimensión tan elevada en longitud no es habitual en las redes empleadas para diagnóstico automático de apea. Como opción se podría dividir la señal SpO_2 en segmentos de tiempo más cortos y generar los espectrogramas de cada uno de estos. Sería posible estimar por segmento el AHI y sumar todos para dar una estimación después para toda la señal. Esto se adaptaría mejor a la red y permitía un mejor análisis temporal.
- Evaluación de arquitecturas más recientes: Se contempla la realización de pruebas adicionales en otras redes efectivas para imagen más avanzadas. Entrenamientos con redes como *EfficientNet*, *Inception* u otros tipos de ResNet podrían ofrecer resultados superiores.

REFERENCIAS

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

Akter, S. *et al.* (2025) *DREAM: A novel explainable neural network for detecting sleep apnea using single-lead ECG signals*. Available at: https://ssrn.com/abstract=5139337.

Al-Beltagi, M. *et al.* (2024) 'Pulse oximetry in pediatric care: Balancing advantages and limitations', *World Journal of Clinical Pediatrics*, 13(3). Available at: https://doi.org/10.5409/wjcp.v13.i3.96950.

Becherer, N. *et al.* (2019) 'Improving optimization of convolutional neural networks through parameter fine-tuning', *Neural Computing and Applications*, 31(8), pp. 3469–3479. Available at: https://doi.org/10.1007/s00521-017-3285-0.

Bengio, Y., Goodfellow, I. and Courville, A. (2015) *Deep Learning*.

Berry, R.B. *et al.* (2018) *The AASM Manual for the Scoring of Sleep and Associated Events RULES, TERMINOLOGY ANO TECHNICAL SPECIFICATIONS VERSION 2.5 AASM Sconng Manual Version 2.5 1*.

Bhatt, D. *et al.* (2021a) 'Cnn variants for computer vision: History, architecture, application, challenges and future scope', *Electronics (Switzerland)*. MDPI. Available at: https://doi.org/10.3390/electronics10202470.

Bhatt, D. *et al.* (2021b) 'Cnn variants for computer vision: History, architecture, application, challenges and future scope', *Electronics (Switzerland)*. MDPI. Available at: https://doi.org/10.3390/electronics10202470.

Chang, S.J. and Chae, K.Y. (2010) 'Obstructive sleep apnea syndrome in children: Epidemiology, pathophysiology, diagnosis and sequelae', *Korean Journal of Pediatrics*. Korean Pediatric Society, pp. 863–871. Available at: https://doi.org/10.3345/kjp.2010.53.10.863.

Childhood Adenotonsillectomy Trial (no date). Available at: https://sleepdata.org/datasets/chat (Accessed: 6 July 2025).

Choudhury, M. *et al.* (2025) 'Explainable AI-driven scalogram analysis and optimized transfer learning for sleep apnea detection with single-lead electrocardiograms', *Computers in Biology and Medicine*, 187. Available at: https://doi.org/10.1016/j.compbiomed.2025.109769.

Crowson, M.G. *et al.* (2023) 'Paediatric sleep apnea event prediction using nasal air pressure and machine learning', *Journal of Sleep Research*, 32(4). Available at: https://doi.org/10.1111/jsr.13851.

Giavarina, D. (2015) ‘Understanding Bland Altman analysis’, *Biochemia Medica*, 25(2), pp. 141–151. Available at: https://doi.org/10.11613/BM.2015.015.

Grandini, M., Bagli, E. and Visani, G. (2020) ‘Metrics for Multi-Class Classification: an Overview’. Available at: http://arxiv.org/abs/2008.05756.

Gulotta, G. *et al.* (2019) ‘Risk factors for obstructive sleep apnea syndrome in children: State of the art’, *International Journal of Environmental Research and Public Health*. MDPI AG. Available at: https://doi.org/10.3390/ijerph16183235.

Gunning, D. *et al.* (2019) ‘XAI-Explainable artificial intelligence’, *Science Robotics*, 4(37). Available at: https://doi.org/10.1126/scirobotics.aay7120.

Gupta, K., Bajaj, V. and Ansari, I.A. (2022) ‘OSACN-Net: Automated Classification of Sleep Apnea Using Deep Learning Model and Smoothed Gabor Spectrograms of ECG Signal’, *IEEE Transactions on Instrumentation and Measurement*, 71. Available at: https://doi.org/10.1109/TIM.2021.3132072.

Gutiérrez-Tobal, G.C. *et al.* (2022) ‘Reliability of machine learning to diagnose pediatric obstructive sleep apnea: Systematic review and meta-analysis’, *Pediatric Pulmonology*, 57(8), pp. 1931–1943. Available at: https://doi.org/10.1002/ppul.25423.

He, K. *et al.* (2016) *Deep Residual Learning for Image Recognition*. Available at: http://image-net.org/challenges/LSVRC/2015/.

Herrero-Tudela, M. *et al.* (2025) ‘An explainable deep-learning model reveals clinical clues in diabetic retinopathy through SHAP’, *Biomedical Signal Processing and Control*, 102. Available at: https://doi.org/10.1016/j.bspc.2024.107328.

Hess, D.R. (2016) ‘Pulse oximetry: Beyond SpO2’, *Respiratory Care*, 61(12), pp. 1671–1680. Available at: https://doi.org/10.4187/respcare.05208.

Hornero, R. *et al.* (2017) ‘Nocturnal oximetry-based evaluation of habitually snoring children’, *American Journal of Respiratory and Critical Care Medicine*, 196(12), pp. 1591–1598. Available at: https://doi.org/10.1164/rccm.201705-0930OC.

Jadon, A., Patil, A. and Jadon, S. (2022) ‘A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting’. Available at: http://arxiv.org/abs/2211.02989.

Janiesch, C., Zschech, P. and Heinrich, K. (2021) ‘Machine learning and deep learning’. Available at: https://doi.org/10.1007/s12525-021-00475-2/Published.

Jiménez-García, J. *et al.* (2022) ‘A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry’, *Computers in Biology and Medicine*, 147. Available at: https://doi.org/10.1016/j.compbiomed.2022.105784.

Jiménez-García, J. *et al.* (2024) ‘An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals’, *Biomedical*

Signal Processing and Control, 87. Available at: https://doi.org/10.1016/j.bspc.2023.105490.

Kaditis, A., Kheirandish-Gozal, L. and Gozal, D. (2015) 'Pediatric OSAS: Oximetry can provide answers when polysomnography is not available', *Sleep Medicine Reviews*. W.B. Saunders Ltd, pp. 96–105. Available at: https://doi.org/10.1016/j.smrv.2015.05.008.

Koo, T.K. and Li, M.Y. (2016) 'A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research', *Journal of Chiropractic Medicine*, 15(2), pp. 155–163. Available at: https://doi.org/10.1016/j.jcm.2016.02.012.

Lecun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*. Nature Publishing Group, pp. 436–444. Available at: https://doi.org/10.1038/nature14539.

Lin, Y. *et al.* (2024) 'Wavelet transform and deep learning-based obstructive sleep apnea detection from single-lead ECG signals', *Physical and Engineering Sciences in Medicine*, 47(1), pp. 119–133. Available at: https://doi.org/10.1007/s13246-023-01346-0.

Linh, T.T.D. *et al.* (2024) 'Detection of preceding sleep apnea using ECG spectrogram during CPAP titration night: A novel machine-learning and bag-of-features framework', *Journal of Sleep Research*, 33(3). Available at: https://doi.org/10.1111/jsr.13991.

Liukkonen, K. *et al.* (2012) 'Symptoms at presentation in children with sleep-related disorders', *International Journal of Pediatric Otorhinolaryngology*, 76(3), pp. 327–333. Available at: https://doi.org/10.1016/j.ijporl.2011.12.002.

Lloberes, P. *et al.* (2011) 'Diagnóstico y tratamiento del síndrome de apneas-hipopneas del sueño', *Archivos de Bronconeumologia*, 47(3), pp. 143–156. Available at: https://doi.org/10.1016/j.arbres.2011.01.001.

Lundberg, S.M., Allen, P.G. and Lee, S.-I. (2017) *A Unified Approach to Interpreting Model Predictions*. Available at: https://github.com/slundberg/shap.

Magalang, U.J. *et al.* (2003) *Prediction of the Apnea-Hypopnea Index From Overnight Pulse Oximetry* Clinical Investigations*. Available at: www.chestjournal.org.

Marcus, C. (2012) *Diagnosis and management of childhood obstructive sleep apnea syndrome*, *PEDIATRICS*. Available at: http://publications.aap.org/pediatrics/article-pdf/130/3/e714/1442504/peds_2012-1672.pdf.

Marcus, C.L. *et al.* (2012) 'Diagnosis and management of childhood obstructive sleep apnea syndrome', *Pediatrics*, 130(3), pp. e714–e755.

Marcus, C.L. *et al.* (2013) 'A Randomized Trial of Adenotonsillectomy for Childhood Sleep Apnea', *New England Journal of Medicine*, 368(25), pp. 2366–2376. Available at: https://doi.org/10.1056/nejmoa1215881.

Mashrur, F.R. *et al.* (2021) 'SCNN: Scalogram-based convolutional neural network to detect obstructive sleep apnea using single-lead electrocardiogram signals', *Computers*

in Biology and Medicine, 134. Available at: https://doi.org/10.1016/j.compbiomed.2021.104532.

Moradhasel, B. *et al.* (2023) 'Spectrogram classification of patient chin electromyography based on deep learning: A novel method for accurate diagnosis obstructive sleep apnea', *Biomedical Signal Processing and Control*, 79. Available at: https://doi.org/10.1016/j.bspc.2022.104215.

Moradi, R., Berangi, R. and Minaei, B. (2020) 'A survey of regularization strategies for deep models', *Artificial Intelligence Review*, 53(6), pp. 3947–3986. Available at: https://doi.org/10.1007/s10462-019-09784-7.

Mortazavi, E. *et al.* (2023) 'Assessing Pediatric Sleep Apnea-Hypopnea Severity: Analyzing SpO2 Signals Spectrograms with Inception V3 Model', in *2023 9th International Conference on Control, Instrumentation and Automation, ICCIA 2023*. Institute of Electrical and Electronics Engineers Inc. Available at: https://doi.org/10.1109/ICCIA61416.2023.10506393.

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

Niroshana, S.M.I. *et al.* (2021) 'A fused-image-based approach to detect obstructive sleep apnea using a single-lead ECG and a 2D convolutional neural network', *PLoS ONE*, 16(4 April). Available at: https://doi.org/10.1371/journal.pone.0250618.

Nitzan, M., Romem, A. and Koppel, R. (2014) 'Pulse oximetry: Fundamentals and technology update', *Medical Devices: Evidence and Research*. Dove Medical Press Ltd, pp. 231–239. Available at: https://doi.org/10.2147/MDER.S47319.

Okumura, S. (2011) 'The Short Time Fourier Transform and Local Signals'.

Roman Jaeschke, Gordon H. Guyatt and David L. Sackett (1994) 'Users' guides to the medical literature: III. How to use an article about a diagnostic test B. What are the results and will they help me in caring for my patients?'

Romero, H.E. *et al.* (2022) 'Acoustic Screening for Obstructive Sleep Apnea in Home Environments Based on Deep Neural Networks', *IEEE Journal of Biomedical and Health Informatics*, 26(7), pp. 2941–2950. Available at: https://doi.org/10.1109/JBHI.2022.3154719.

Salih, A.M. *et al.* (2024) 'A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME', *Advanced Intelligent Systems* [Preprint]. Available at: https://doi.org/10.1002/aisy.202400304.

Salih, A.M. *et al.* (2025) 'A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME', *Advanced Intelligent Systems*, 7(1). Available at: https://doi.org/10.1002/aisy.202400304.

Samek, W. *et al.* (eds.) (2019) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Cham: Springer International Publishing (Lecture Notes in Computer Science). Available at: https://doi.org/10.1007/978-3-030-28954-6.

Song, Y. *et al.* (2023) 'AHI estimation of OSAHS patients based on snoring classification and fusion model', *American Journal of Otolaryngology - Head and Neck Medicine and Surgery*, 44(5). Available at: https://doi.org/10.1016/j.amjoto.2023.103964.

Stankovic, L. and Dakovic, M. (2013) *Time-frequency signal analysis with applications*. Available at: https://www.researchgate.net/publication/260061447.

Stewart, M. *et al.* (2023) 'Decision Fusion in Automated Sleep Apnea Classification Using Multple Polysomnography Sensors and Convolutional Neural Networks', in *2023 IEEE Sensors Applications Symposium, SAS 2023 - Proceedings*. Institute of Electrical and Electronics Engineers Inc. Available at: https://doi.org/10.1109/SAS58821.2023.10254145.

Sun, R.Y. (2020) 'Optimization for Deep Learning: An Overview', *Journal of the Operations Research Society of China*, 8(2), pp. 249–294. Available at: https://doi.org/10.1007/s40305-020-00309-6.

Tanci, K. and Hekim, M. (2025) 'Classification of sleep apnea syndrome using the spectrograms of EEG signals and YOLOv8 deep learning model', *PeerJ Computer Science*, 11. Available at: https://doi.org/10.7717/peerj-cs.2718.

Tauman, R. and Gozal, D. (2011) 'Obstructive sleep apnea syndrome in children', *Expert Review of Respiratory Medicine*. Expert Reviews Ltd., pp. 425–440. Available at: https://doi.org/10.1586/ers.11.7.

Telea, A. (2004) *An Image Inpainting Technique Based on the Fast Marching Method*. APA. Available at: http://www.rug.nl/research/portal.

Tyagi, P.K. and Agarwal, D. (2023) 'Systematic review of automated sleep apnea detection based on physiological signal data using deep learning algorithm: a meta-analysis approach', *Biomedical Engineering Letters*. Springer Verlag, pp. 293–312. Available at: https://doi.org/10.1007/s13534-023-00297-5.

Ullah, N. *et al.* (2023) 'DCDA-Net: Dual-convolutional dual-attention network for obstructive sleep apnea diagnosis from single-lead electrocardiograms', *Engineering Applications of Artificial Intelligence*, 123. Available at: https://doi.org/10.1016/j.engappai.2023.106451.

Vaquerizo-Villar, F. *et al.* (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), pp. 2906–2916. Available at: https://doi.org/10.1109/JBHI.2020.3048901.

Vihinen, M. (2012) ‘How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis.’, *BMC genomics*, 13 Suppl 4. Available at: https://doi.org/10.1186/1471-2164-13-S4-S2.

Wacker, M. and Witte, H. (2013) ‘Time-frequency techniques in biomedical signal analysis: A tutorial review of similarities and differences’, *Methods of Information in Medicine*, 52(4), pp. 279–296. Available at: https://doi.org/10.3414/ME12-01-0083.

Wali, S. *et al.* (2020) ‘The correlation between oxygen saturation indices and the standard obstructive sleep apnea severity’, *Annals of Thoracic Medicine*, 15(2), pp. 70–75. Available at: https://doi.org/10.4103/atm.ATM_215_19.

Wang, B. *et al.* (2022) ‘Obstructive Sleep Apnea Detection Based on Sleep Sounds via Deep Learning’, *Nature and Science of Sleep*, 14, pp. 2033–2045. Available at: https://doi.org/10.2147/NSS.S373367.

Wang, E., Koprinska, I. and Jeffries, B. (2023) ‘Sleep Apnea Prediction Using Deep Learning’, *IEEE Journal of Biomedical and Health Informatics*, 27(11), pp. 5644–5654. Available at: https://doi.org/10.1109/JBHI.2023.3305980.

Weir, J.P. (2005) ‘Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM’, *Journal of Strength and Conditioning Research*, pp. 231–240. Available at: https://doi.org/10.1519/15184.1.

Wu, J. (2017) *Introduction to Convolutional Neural Networks*.

Wu, Y. *et al.* (2021) ‘A novel approach to diagnose sleep apnea using enhanced frequency extraction network’, *Computer Methods and Programs in Biomedicine*, 206. Available at: https://doi.org/10.1016/j.cmpb.2021.106119.

Xu, W., Fu, Y.L. and Zhu, D. (2023) ‘ResNet and its application to medical image processing: Research progress and challenges’, *Computer Methods and Programs in Biomedicine*, 240. Available at: https://doi.org/10.1016/j.cmpb.2023.107660.

Yu, Y. *et al.* (2021) ‘FASSNet: Fast apnea syndrome screening neural network based on single-lead electrocardiogram for wearable devices’, *Physiological Measurement*, 42(8). Available at: https://doi.org/10.1088/1361-6579/ac184e.

Zhou, Y. and Kang, K. (2024) ‘Multi-Feature Automatic Extraction for Detecting Obstructive Sleep Apnea Based on Single-Lead Electrocardiography Signals’, *Sensors*, 24(4). Available at: https://doi.org/10.3390/s24041159.