Inteligencia artificial explicable basada en *Shapley values* para interpretar modelos de *deep learning* orientados a detectar la apnea del sueño pediátrica usando flujo aéreo y oximetría

J. Jiménez García[1,2], G.C. Gutiérrez Tobal[1,2], M. García Gadañón[1,2], J. Michiels[3], G. Gagliardi[3,4], D. Gozal[5], M. de Vos[3,6], R. Hornero[1,2]

[1] Grupo de Ingeniería Biomédica, Universidad de Valladolid, Valladolid, España, jorge.jimenez@gib.tel.uva.es

[2] CIBER de Bioingeniería, Biomateriales y Nanomedicina, Instituto de Salud Carlos III, Valladolid, España

[3] STADIUS Center for Dynamical Systems, Signal Processing, and Data Analytics, ESAT, KU Leuven, Leuven, Belgium

[4] Department of Information Engineering, University of Florence, Florence, Italy

[5] Office of The Dean, Joan C. Edwards School of Medicine, Marshall University, Huntington, WV, USA

[6] Department of Development and Regeneration, Faculty of Medicine, KU Leuven, Leuven, Belgium

Resumen

El trastorno de la apnea obstructiva del sueño (AOS) afecta al 5% de los niños y provoca complicaciones relacionadas con el desarrollo en la edad infantil. Recientemente, se han propuesto técnicas de análisis de señales polisomnográficas nocturnas basadas en deep learning (DL) para simplificar el diagnóstico de la AOS. Sin embargo, la opacidad de estos algoritmos y la dificultad para interpretar los patrones extraídos de dichas señales han lastrado la viabilidad de estas soluciones. El objetivo de este estudio fue analizar una técnica de inteligencia artificial explicable (XAI) basada en Shapley Additive Explanations (SHAP) para interpretar los modelos de DL y las inferencias que realizan sobre las señales de flujo aéreo (FA) y oximetría (SpO_2). Se implementó la técnica WindowSHAP para explicar la contribución de ventanas cortas de las señales de FA y SpO_2 a la estimación del modelo. Se empleó una base de datos pública con 1.638 registros nocturnos para obtener los modelos DL y el algoritmo WindowSHAP. Los valores SHAP más elevados se identificaron en segmentos cortos que contenían desaturaciones en la SpO_2 y reducciones de la amplitud del FA, eventos relacionados con la presencia de eventos de apnea/hipopnea. La relación directa entre los patrones más relevantes de ambas señales y las atribuciones obtenidas mediante WindowSHAP reforzaron la explicabilidad de unos modelos de DL precisos, pero hasta ahora opacos. En resumen, la técnica de XAI WindowSHAP contribuyó satisfactoriamente a mejorar la credibilidad e interpretabilidad de los métodos automáticos de ayuda al diagnóstico de la AOS pediátrica.

1. Introducción

La apnea obstructiva del sueño (AOS) es uno de los trastornos del sueño más frecuentes en la población infantil, con una prevalencia global estimada entre el 1% y el 5% [1]. Los episodios recurrentes de cese total o parcial de la respiración (apneas e hipopneas) durante el sueño, característicos de la AOS, provocan la interrupción de las fases del sueño y alteran su desarrollo normal [1]. El diagnóstico y tratamiento temprano de la AOS puede minimizar sus consecuencias negativas, que abarcan desde déficits neurocognitivos y trastornos conductuales hasta complicaciones cardio-metabólicas y problemas de crecimiento [1]. La prueba diagnóstica estándar es la polisomnografía nocturna (PSG), que consiste en el registro y análisis de numerosas señales (cardiorrespiratorias, neuronales, musculares, etc.) durante la noche en una unidad del sueño especializada [1]. El principal parámetro calculado a partir de la PSG para diagnosticar la AOS y su nivel de severidad es el índice de apnea/hipopnea (IAH), la tasa de episodios de apnea e hipopnea por cada hora de sueño (e/h) [1]. Sin embargo, la PSG es una prueba complicada, costosa y la disponibilidad de instalaciones hospitalarias es limitada [1]. Estos inconvenientes motivan la búsqueda de diagnósticos más simples y baratos, como pueden ser el registro de un menor número de señales y/o automatizar su análisis [2], [3].

Las técnicas de aprendizaje profundo (*deep learning*, DL) han demostrado una elevada utilidad diagnóstica en el contexto de la AOS pediátrica, incluso reduciendo al máximo las señales involucradas [4]–[6]. Las señales de flujo aéreo (FA) y oximetría (saturación de oxígeno en sangre, SpO_2) son los registros nocturnos más relacionados con las definiciones de los eventos de apnea/hipopnea y las desaturaciones de oxígeno asociadas [7]. Estas dos señales se han utilizado recientemente para desarrollar métodos de DL con el objetivo de ayudar en el diagnóstico de la AOS en niños [5], [6]. Sin embargo, los mecanismos internos de las redes de DL más consolidadas, como las redes neuronales convolucionales (*Convolutional Neural Network*, CNN) y recurrentes (*Recurrent Neural Network*, RNN), son difícilmente interpretables y los convierten en algoritmos opacos [8]. La inteligencia artificial explicable (*eXplainable Artificial Intelligence*, XAI) es un campo que se enfoca en proporcionar transparencia a los algoritmos de aprendizaje automático, especialmente las redes de DL [8]. Sin embargo, los métodos de XAI no se han aplicado de manera extensiva para la ayuda al diagnóstico de la AOS. Hasta el momento, solo se ha aplicado la técnica *Gradient-Weighted Class Activation Mapping* (Grad-CAM), restringida al análisis de redes CNN, para interpretar las predicciones realizadas mediante *heatmaps* sobre las señales [6]. La técnica *Shapley Additive Explanations* (SHAP) es una de las más extendidas y es aplicable a cualquier modelo (*model agnostic*), pero con limitaciones en cuanto al análisis de señales [9], [10]. El objetivo de este

ISBN: 978-84-09-67332-2

estudio fue analizar una técnica derivada de SHAP, especialmente diseñada para el análisis de series temporales, y aplicada sobre las señales de FA y SpO_2.

2. Materiales y métodos

2.1. Sujetos y señales

En este trabajo se utilizó la base de datos pública *Childhood Adenotonsilletomy Trial* (CHAT), un estudio aleatorizado que evaluó la efectividad de un tratamiento quirúrgico de la AOS infantil frente a una alternativa más conservadora [11]. Se emplearon 1.638 registros nocturnos de FA y SpO_2 procedentes de PSG realizadas a niños entre 5 y 10 años, con y sin AOS, divididos aleatoriamente en 3 grupos: entrenamiento, validación y test (Tabla 1). El primer conjunto se empleó para desarrollar los algoritmos de DL y XAI, mientras que el segundo se utilizó para optimizar hiperparámetros. El conjunto de test sirvió para evaluar la utilidad diagnóstica del modelo de DL, así como analizar los resultados obtenidos por el algoritmo basado en SHAP. Todos los registros contaban con anotaciones de la posición de los eventos de apnea, hipopnea y las desaturaciones, entre otros.

Las señales de FA y SpO_2 se preprocesaron con el objetivo de homogeneizar su frecuencia de muestreo y su amplitud. Ambos registros nocturnos se muestrearon a la misma tasa (fs=10 Hz). Se atenuó el ruido de las señales de FA usando un filtro paso-bajo entre 0 y 1,5 Hz y se normalizó la amplitud del FA adaptativamente como en estudios previos [5]. Las señales se dividieron en épocas de 5 minutos (3.000 muestras) y en segmentos de 30 minutos (18.000 muestras), para emplearlas como entradas de los modelos de DL [5], [6]. La longitud de las épocas y de los segmentos se escogió para lograr que los algoritmos de DL modelaran la recurrencia de los eventos apneicos y sus desaturaciones, que normalmente se concentran en clústeres de varios minutos. Las épocas y segmentos se etiquetaron con el número de apneas/hipopneas presentes en cada una.

2.2. Algoritmos de *deep learning*: CNN y CNN+RNN

Los modelos de DL elegidos fueron dos algoritmos desarrollados y evaluados en estudios previos empleando las señales de FA y SpO_2 [5], [6]. El primero es una CNN orientada a detectar la AOS pediátrica a partir de estos dos registros. Esta red toma como entrada épocas de 5 min de ambas señales, que se procesan en bloques consecutivos formados por capas convolucionales, *batch normalization*, *max pooling* y *dropout* [5]. La salida de la CNN es una capa *dense* que computa el número de eventos detectados en las señales. El segundo algoritmo es una combinación de la CNN anterior con una RNN de tipo *Gated Recurrent Unit* (GRU) mediante *transfer learning* (CNN+RNN) [6]. Los segmentos de 30 min se distribuyeron en 6 épocas de 5 min, que son procesadas en bloques *Time Distributed* (TD) que encapsulan las capas de la CNN descritas anteriormente. La capa GRU analiza las secuencias generadas en los bloques TD e infiere la cantidad de eventos presentes en cada secuencia [6]. Ambas arquitecturas estaban orientadas a estimar el IAH a partir de los eventos de apnea/hipopnea detectados. La configuración de hiperparámetros en los modelos CNN y CNN+RNN se optimizó utilizando los datos del conjunto de validación, como se indica en estudios previos [5], [6]. Ambos modelos de DL demostraron un rendimiento diagnóstico elevado en dichos estudios, con precisiones entre el 84,64% (1 e/h) y el 94,44% (10 e/h) para la red CNN y entre el 87,25% (1 e/h) y el 93,46% utilizando la arquitectura CNN+RNN [5], [6].

	Entrenam.	Validación	Test
Sujetos (n)	1006	326	306
Edad (años)	7 [6; 8]	7 [6; 8]	6,9 [6; 8]
Varones (%)	46,8%	47,9%	43,8%
IMC (kg/m^2)	17,4 [15,6; 21,7]	17,1 [15,4; 21,8]	17,6 [15,7; 21,7]
IAH (eventos/h)	2,6 [1,1; 5,9]	2,4 [1,2; 5,8]	2,3 [1,1; 6,2]
IAH<1 (n)	219	69	67
1≤IAH<5 (n)	496	168	148
5≤IAH<10 (n)	160	44	49
IAH≥10 (n)	131	45	42

Tabla 1. *Datos sociodemográficos y clínicos de los sujetos.*

2.3. *Explainable artificial intelligence*: WidowSHAP

La etapa de análisis e interpretación de las predicciones producidas por los modelos de DL se llevó a cabo mediante la técnica WindowSHAP [10]. Los métodos de la familia SHAP provienen de la teoría de juegos, en los que los valores de Shapley tratan de repartir la recompensa de una tarea entre los "jugadores" que han intervenido en ella. De esta manera, el valor de Shapley (ϕ_i) se entiende como la contribución de un participante al resultado [9]. SHAP es un método que se puede aplicar a cualquier modelo de *machine learning*, en el que la recompensa se asimila a la predicción del modelo $\hat{y} = f(\boldsymbol{x})$ y los actores serían las variables de entrada al mismo [9]:

$$f(\boldsymbol{x}) = \phi_0 + \sum_{i=1}^{D} \phi_i, \quad (1)$$

donde el valor $\phi_0 = E[f(\boldsymbol{x})]$ es el valor esperado a la salida si se omitieran todas las variables de entrada. Sin embargo, los modelos de DL que procesan señales en crudo pueden tener un elevado número de muestras como variables de entrada, existiendo mucha redundancia entre las mismas, y SHAP tiene una complejidad algorítmica elevada [10]. Esto hace que una implementación típica como KernelSHAP no sea adecuada para explicar series temporales. WindowSHAP es una adaptación de SHAP para explicar modelos de DL basados en series temporales. Esta técnica divide las señales de entrada en ventanas temporales cortas y estima los valores de Shapley atribuibles a cada ventana mediante KernelSHAP [10]. Las principales ventajas de WindowSHAP frente a SHAP son: (*i*) las explicaciones basadas en pequeñas ventanas son más inteligibles, y (*ii*) el coste computacional se reduce al explicar ventanas en vez de muestras temporales [10].

En este estudio se emplearon ventanas de 10 segundos para obtener las contribuciones de las señales de FA y SpO_2 a las predicciones de los modelos CNN y CNN+RNN mediante WindowSHAP. Esta longitud de ventana está directamente relacionada con la duración típica de los

eventos de apnea/hipopnea (mínimo 2 ciclos respiratorios) [7]. Se emplearon señales del conjunto de entrenamiento como *background* para estimar los valores de SHAP.

2.4. Análisis de resultados

Los valores de Shapley obtenidos se representaron gráficamente sobre las señales de FA y SpO_2 empleando una escala de colores, en la que las contribuciones positivas (partes de la señal que hacen que la predicción crezca) aparecen en color magenta y las negativas (aquellas que hacen que la estimación baje) aparecen en color azul. Se compararon los valores de Shapley correspondientes a ventanas que contenían apneas o hipopneas (marcadas por los expertos sobre la señal de FA) y desaturaciones (marcadas sobre la señal de SpO_2) frente a los tramos de respiración normal mediante *boxplots*. Para el análisis estadístico de los valores de Shapley se emplearon tests no paramétricos de rango de Wilcoxon.

3. Resultados

Los ejemplos mostrados en las Figuras 1 y 2 corresponden a épocas de 5 min de las señales de FA y SpO_2, junto a los valores de Shapley obtenidos al analizar con WindowSHAP el modelo CNN. El tramo de respiración normal (Figura 1) muestra varias ventanas con atribuciones ligeramente negativas y casi nulas, provocando que la predicción de eventos sea cercana a cero. Las ventanas con atribuciones positivas en la Figura 2 resaltan principalmente el descenso momentáneo de la SpO_2 y los cambios de amplitud en el FA. En ambas Figuras se puede apreciar que gran parte de las ventanas temporales sin apneas o hipopneas tienen asociado un valor de Shapley nulo, que no señalaría patrones relevantes. La Figura 3 muestra un tramo de 30 min en el que WindowSHAP atribuyó la salida del modelo CNN+RNN a una agrupación de varios eventos de apnea/hipopnea consecutivos, destacando la contribución de las desaturaciones sobre las fluctuaciones del FA. Al igual que en el modelo CNN, se aprecia que WindowSHAP asigna a las desaturaciones los valores absolutos de Shapley más elevados.

Las atribuciones obtenidas sobre las señales de FA y SpO_2 mediante WindowSHAP en los registros del conjunto de test fueron diferentes en función del tipo de evento y la señal de entrada. Los *boxplots* de la Figura 4 muestran las distribuciones de los valores de WindowSHAP sobre la red CNN, diferenciando entre ventanas sin eventos (normal) frente a ventanas con apneas/hipopneas o desaturaciones. Se aprecian diferencias en todas las comparaciones (p<0,01 tras corrección de Bonferroni), siendo más evidentes cuando se comparan los datos obtenidos en presencia de eventos frente a los tramos normales. También hay que mencionar que los valores de Shapley asociados a las desaturaciones son más elevados que los obtenidos en zonas con apneas/hipopneas en la señal de FA. Esta tendencia también se aprecia al comparar las atribuciones de las señales de FA y SpO_2 en las ventanas temporales en las que coinciden ambos tipos de eventos respiratorios (p<0,01). Por último, los resultados obtenidos con el modelo CNN+RNN fueron totalmente análogos a los logrados con la red CNN. En ambos casos, las diferencias entre los valores de Shapley en ventanas con

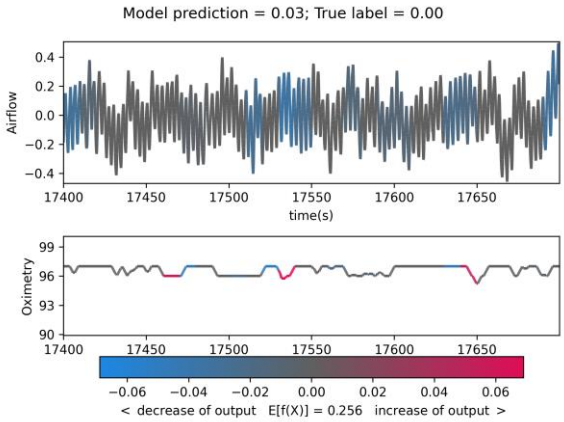

Figura 1.*Atribuciones de WindowSHAP sobre una época de respiración normal empleando el modelo CNN.*

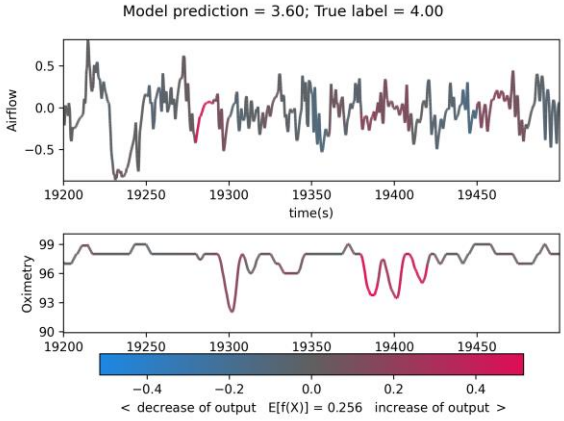

Figura 2.*Atribuciones de WindowSHAP sobre una época con eventos apneicos empleando el modelo CNN.*

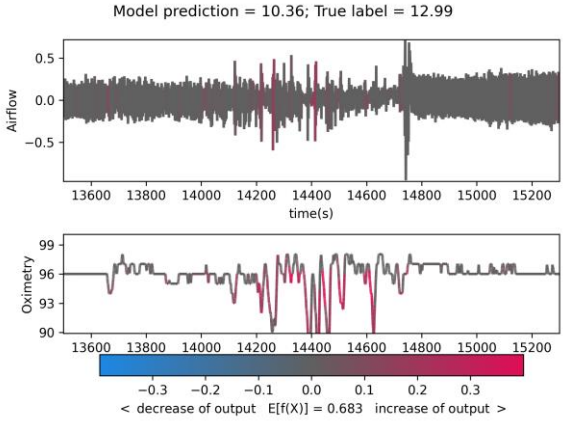

Figura 3. *Atribuciones de WindowsSHAP sobre un segmento que contiene una sucesión de eventos de apnea/hipopnea, empleando la CNN+RNN.*

eventos fueron significativamente mayores que en los tramos normales, por ello no se mostraron explícitamente.

4. Discusión y conclusiones

En este estudio se ha aplicado por primera vez el algoritmo de XAI WindowSHAP para analizar modelos de DL orientados a la detección de la AOS pediátrica. Los únicos precedentes relacionados con la temática de este estudio emplearon Grad-CAM para analizar sus modelos [6], [12].

ISBN: 978-84-09-67332-2

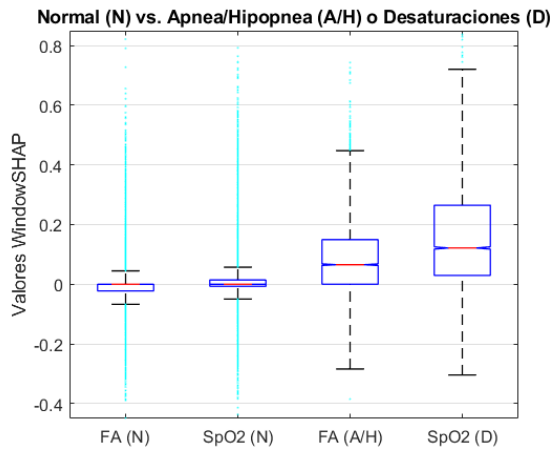

Figura 4.*Boxplots de los valores derivados de WindowSHAP en ventanas de FA y SpO2 de respiración normal, (hipo)apneas y desaturaciones.*

En estos estudios, los *heatmaps* obtenidos se analizaron subjetivamente, y no se profundizó en analizar cuantitativamente la relación entre las explicaciones y las anotaciones de los expertos. WindowSHAP permitió identificar los eventos más importantes en cada señal para detectar la AOS, según los modelos empleados. Además, se pudo cuantificar la contribución de los eventos respiratorios detectados por estos modelos a la predicción. Los resultados sugieren que ambas señales contribuyeron a identificar los eventos, pero la SpO_2 sería más explicativa que el FA, ya que los valores de Shapley son mayores en los tramos con desaturaciones. Esta afirmación estaría en línea con los resultados obtenidos en estudios previos, en los que la capacidad diagnóstica de modelos basados solo en SpO_2 es ligeramente inferior a aquellos que emplearon dos señales [4], [5]. Las principales limitaciones de la metodología presentada en este estudio son: (*i*) a pesar de que CHAT es una base de datos multicentro, se deberían validar los resultados en otras poblaciones; (*ii*) los modelos de DL y XAI probados se han orientado a la detección de la AOS pediátrica, por lo que sería deseable probarlos en otros contextos para verificar su utilidad diagnóstica.

Como conclusión, el método WindowSHAP ayudó a interpretar modelos de DL dirigidos a simplificar el diagnóstico de la AOS pediátrica, permitiendo verificar su funcionamiento e incrementando así su credibilidad.

Agradecimientos

Esta investigación ha sido co-financiada por MICIU/AEI/ 10.13039/501100011033, Unión Europea, FEDER, NextGenerationEU/PRTR (PID2020-115468RB-I00, CPP2022-009735), Programa de Cooperación Transfronteriza Interreg VI-A España-Portugal (POCTEP) 2021-2027 (0043_NET4SLEEP_2_E), por CIBER-BBN (CB19/01/00012) mediante el Instituto de Salud Carlos III, y por *Flanders AI Research Programme*. J. Jiménez García contó con un contrato predoctoral de la Universidad de Valladolid y una ayuda "Movilidad de estudiantes de Doctorado UVa 2023".

Referencias

[1] E. Dehlink *et al.*, "Update on paediatric obstructive sleep apnoea," *J Thorac Dis*, vol. 8, no. 2, pp. 224–235, 2016, doi: 10.3978/j.issn.2072-1439.2015.12.04.

[2] D. Bertoni *et al.*, "Towards Patient-centered Diagnosis of Pediatric Obstructive Sleep Apnea—A Review of Biomedical Engineering Strategies," *Expert Rev. Med. Devices*, vol. 16, no. 7, pp. 617–629, 2019, doi: 10.1080/17434440.2019.1626233.

[3] G. C. Gutiérrez-Tobal *et al.*, "Reliability of machine learning to diagnose pediatric obstructive sleep apnea: Systematic review and meta-analysis," *Pediatr. Pulmonol.*, vol. 57, no. 8, pp. 1931–1943, Aug. 2022, doi: 10.1002/ppul.25423.

[4] F. Vaquerizo-Villar *et al.*, "A Convolutional Neural Network Architecture to Enhance Oximetry Ability to Diagnose Pediatric Obstructive Sleep Apnea," *IEEE J. Biomed. Heal. Informatics*, vol. 25, no. 8, pp. 2906–2916, Aug. 2021, doi: 10.1109/JBHI.2020.3048901.

[5] J. Jiménez-García *et al.*, "A 2D convolutional neural network to detect sleep apnea in children using airflow and oximetry," *Comput. Biol. Med.*, vol. 147, no. August, p. 105784, Aug. 2022, doi: 10.1016/j.compbiomed.2022.105784.

[6] J. Jiménez-García *et al.*, "An explainable deep-learning architecture for pediatric sleep apnea identification from overnight airflow and oximetry signals," *Biomed. Signal Process. Control*, vol. 87, no. September 2023, p. 105490, Jan. 2024, doi: 10.1016/j.bspc.2023.105490.

[7] R. B. Berry *et al.*, "Rules for Scoring Respiratory Events in Sleep: Update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events," *J. Clin. Sleep Med.*, vol. 8, no. 5, pp. 597–619, Oct. 2012, doi: 10.5664/jcsm.2172.

[8] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, no. October 2019, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[9] S. M. Lundberg *et al.*, "A Unified Approach to Interpreting Model Predictions," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[10] A. Nayebi *et al.*, "WindowSHAP: An efficient framework for explaining time-series classifiers based on Shapley values," *J. Biomed. Inform.*, vol. 144, no. June, p. 104438, 2023, doi: 10.1016/j.jbi.2023.104438.

[11] C. L. Marcus *et al.*, "A Randomized Trial of Adenotonsillectomy for Childhood Sleep Apnea," *N. Engl. J. Med.*, vol. 368, no. 25, pp. 2366–2376, 2013, doi: 10.1056/nejmoa1215881.

[12] Á. Serrano Alarcón *et al.*, "Obstructive sleep apnea event detection using explainable deep learning models for a portable monitor," *Front. Neurosci.*, vol. 17, no. July, pp. 1–19, 2023, doi: 10.3389/fnins.2023.1155900.