Clasificación explicable de EEG mediante descriptores compactos y reglas difusas lingüísticas

Javier Aguirre-Eraso[1 [0009−0001−0382−0800]], Carlos Lopez-Molina[1,2 [0000−0002−0904−9834]] y Carmen Vidaurre[3,4,5 [0000−0003−3740−049X]]

[1] Universidad Pública de Navarra (UPNA), Pamplona, España
javier.aguirreeraso@unavarra.es
[2] IdiSNA, Navarra Institute for Health Research, NavarraBiomed, Pamplona, España
[3] Basque Center on Cognition, Brain and Language (BCBL), Donostia-San Sebastián, España
[4] Ikerbasque, Bilbao, España
[5] TU-Berlin / BIFOLD Center, Berlín, Alemania

Resumen El aprendizaje profundo se ha posicionado como la solución predeterminada en la clasificación de señales biomédicas. Sin embargo, su opacidad limita su aplicabilidad en escenarios donde las predicciones requieren auditoría o justificación. Este estudio evalúa el equilibrio entre rendimiento y explicabilidad en una tarea de clasificación de edad mediante EEG de reposo (conjunto LEMON, N=203). Comparamos el rendimiento de una red neuronal convolucional (CNN) sobre matrices espectrales complejas frente a un framework difuso basado en descriptores compactos de conectividad biológicamente interpretables. La validación cruzada muestra una diferencia predictiva marginal: la CNN obtiene un accuracy de 0.729 (AUC: 0.800) frente al 0.709 (AUC: 0.776) del modelo difuso. A cambio de esta mínima diferencia, el modelo transparente justifica sus decisiones explícitamente utilizando una base de solo cinco reglas lingüísticas legibles. Estos resultados sugieren que, en ciertos problemas, compensa no perseguir el máximo rendimiento predictivo absoluto si a cambio se prioriza la interpretabilidad y la trazabilidad del modelo.

Palabras clave: Inteligencia artificial explicable · EEG · reglas difusas · conectividad funcional · aprendizaje profundo

1. Introducción

En los últimos años, el deep learning se ha consolidado como la primera opción para abordar muchos problemas de clasificación. La razón es comprensible: cuando hay suficientes datos, las redes neuronales pueden aprender representaciones complejas sin exigir un diseño manual de variables. En EEG, por ejemplo, las arquitecturas convolucionales se han convertido en una de las familias más empleadas por su capacidad para explotar estructura espacial y espectral, y por su flexibilidad frente a distintos paradigmas experimentales [10,7]. Esta tendencia,

sin embargo, ha creado una inercia metodológica peligrosa: asumir que un modelo más profundo o complejo es necesariamente la mejor solución.

La interpretabilidad no es un requisito estético. En salud, una predicción puede influir en decisiones clínicas o en la generación de hipótesis sobre mecanismos fisiológicos. En finanzas, un modelo puede afectar a la concesión de crédito, la detección de fraude o la gestión de riesgo, escenarios donde la trazabilidad es inseparable de la responsabilidad. En neurotecnología, además, el modelo no solo clasifica: también puede orientar qué patrones cerebrales merecen ser estudiados. En estos dominios, explicar los resultados de una caja negra puede quedarse corto si el objetivo no es solo acertar, sino justificar y discutir la decisión. Rudin ha defendido que, cuando hay consecuencias relevantes o necesidad de auditoría, conviene considerar modelos intrínsecamente interpretables antes de recurrir a cajas negras explicadas *a posteriori* [11]. En la misma línea, la literatura sobre inteligencia artificial explicable (XAI) subraya que transparencia, responsabilidad y confianza dependen de que las explicaciones sean comprensibles para los usuarios del sistema, no solo para quien lo programó [1]. El punto central, por tanto, no es negar la utilidad del deep learning, sino preguntarse cuándo compensa usarlo. Si un modelo interpretable obtiene resultados muy próximos a los de una red neuronal, el pequeño margen predictivo de la red puede no justificar la pérdida de legibilidad. Esta cuestión es especialmente relevante en muestras moderadas, donde las arquitecturas flexibles pueden aprender patrones útiles, pero también explotar regularidades difíciles de auditar. Frente a ello, un modelo basado en descriptores comprensibles y reglas lingüísticas ofrece una alternativa mucho más clara en su razonamiento.

El EEG de reposo ofrece un escenario adecuado para estudiar este equilibrio. Las señales registradas reflejan interacciones distribuidas entre regiones cerebrales y bandas de frecuencia, y los cambios asociados al envejecimiento sano pueden manifestarse como variaciones sutiles de potencia, sincronía o estructura de conectividad [6]. Además, las matrices espectrales complejas contienen información que permite construir descriptores biológicos orientados a interacciones no instantáneas entre diferentes regiones cerebrales. En particular, la parte imaginaria de la coherencia se propuso para reducir la influencia del efecto de conducción de volumen [9].

En este trabajo se comparan de forma controlada dos aproximaciones sobre la misma tarea binaria, distinguir participantes jóvenes y mayores en el conjunto LEMON [2]. La primera es una CNN que opera directamente sobre las matrices espectrales complejas. La segunda es ExFuzzy, un framework que combina selección de descriptores compactos, inspirados en propiedades biológicas de la conectividad, con reglas difusas lingüísticas optimizadas evolutivamente [3]. La contribución principal no es reclamar un nuevo estado del arte, sino mostrar que una solución interpretable puede acercarse mucho al rendimiento de una CNN y, al mismo tiempo, devolver reglas comprensibles sobre las variables que sostienen la clasificación.

2. Materiales y métodos

2.1. Datos y representación de entrada

Se empleó la condición de ojos abiertos (EO) del conjunto Leipzig Mind-Brain-Body, una base multimodal diseñada para estudiar relaciones entre cerebro, cognición, emoción y fisiología periférica en adultos jóvenes y mayores [2]. El conjunto de datos contiene 203 participantes: 138 etiquetados como *Young* y 65 como *Old*. Para cada participante se dispone de matrices espectrales complejas en cinco bandas, Delta, Theta, Alpha, Beta y Gamma, con 61 canales por banda.

La representación de partida es la misma para ambos enfoques. Para la CNN se construyó un tensor concatenando las partes real e imaginaria de las cinco matrices espectrales, con tamaño final $10 \times 61 \times 61$. Para el modelo difuso, cada matriz compleja se resumió mediante descriptores compactos de conectividad. En cada banda se calculó fuerza media, entropía de fase ponderada por magnitud y heterogeneidad de centralidad; además, se añadió una medida de interferencia entre Alpha y Delta. Esta compresión reduce la dimensionalidad y facilita la interpretación, pero también elimina información espacial fina que la CNN sí conserva.

2.2. Clasificador convolucional

La CNN se diseñó como un baseline deliberadamente pequeño, ajustado al tamaño de muestra disponible. La arquitectura recibe 10 canales de entrada y contiene dos bloques convolucionales: una convolución 5×5 con 16 filtros, activación ReLU y *max pooling*, seguida de una convolución 3×3 con 32 filtros, activación ReLU, otro *max pooling* y *adaptive average pooling*. La cabeza predictora aplana la representación y aplica una capa densa de 16 unidades con ReLU antes de la salida binaria. El entrenamiento utilizó *BCEWithLogitsLoss* ponderada por la proporción de clases del pliegue de entrenamiento, Adam con tasa de aprendizaje 10^{-3}, *weight decay* 10^{-4}, lotes de 16 muestras, un máximo de 60 épocas y *early stopping* con paciencia de 8 épocas. En cada pliegue externo, el 20 % del entrenamiento se reservó como validación interna estratificada y la normalización por canal se estimó solo con las muestras de entrenamiento.

El umbral de decisión de la CNN se mantuvo fijo en 0.5. No se optimizó por pliegue porque la CNN se usa como referencia compacta, no como techo de rendimiento exhaustivamente ajustado. Esta distinción es importante porque el objetivo experimental es comparar rendimiento y transparencia bajo un presupuesto razonable, no agotar la búsqueda de arquitecturas neuronales.

2.3. Clasificador difuso explicable

El modelo transparente se implementó con ExFuzzy, una biblioteca orientada a inteligencia artificial explicable simbólica mediante programación difusa [3]. En cada fold, la selección de variables se ajustó únicamente con las muestras de entrenamiento mediante *SelectKBest* y estadístico ANOVA F, manteniendo

$k = 5$ descriptores. Las variables seleccionadas se escalaron al intervalo $[0, 1]$ y se aplicó Edited Nearest Neighbours sobre el entrenamiento como limpieza local de muestras. El clasificador final usó particiones difusas de tipo 1 con tres términos lingüísticos, *Low*, *Medium* y *High*, y una base de reglas con un máximo de seis reglas y hasta dos antecedentes por regla.

Las reglas se optimizaron con el backend evolutivo de ExFuzzy basado en *pymoo*. Este backend realiza una búsqueda poblacional de bases de reglas candidatas, variando antecedentes y consecuentes lingüísticos y seleccionando las soluciones de acuerdo con la función de ajuste interna de ExFuzzy, siguiendo el enfoque general de los sistemas difusos evolutivos [5]. La configuración empleó 40 generaciones, población de 22 individuos, probabilidad de variación de 0.20, η de cruce SBX igual a 5.0, η de mutación igual a 10.0 y torneo de tamaño 2. En cada fold se entrenaron tres modelos con semillas 42, 43 y 44 y se promediaron sus probabilidades. El umbral para la clase *Old* se fijó en 0.77 a partir de la exploración previa y se mantuvo constante durante la validación externa. Las reglas finales se reajustaron con todos los datos y la configuración seleccionada, únicamente para describir el modelo lingüístico de referencia; las métricas reportadas proceden de predicciones fuera de muestra en los folds externos.

2.4. Evaluación

Ambos modelos se evaluaron mediante validación cruzada estratificada de cinco particiones. Se reportaron accuracy, área bajo la curva ROC (AUC), coeficiente de correlación de Matthews (MCC), balanced accuracy, F1 macro y recall por clase. Estas métricas permiten resumir tanto el rendimiento global como el comportamiento por clase en una muestra con tamaños de grupo distintos. La Tabla 1 resume las diferencias principales entre las dos configuraciones antes de presentar los resultados.

Tabla 1. Configuración resumida de los dos modelos comparados.

Componente	CNN compacta	ExFuzzy
Entrada	Tensor real/imaginario $10 \times 61 \times 61$	16 descriptores iniciales de conectividad
Selección de variables	No aplicada	ANOVA, $k = 5$
Preprocesamiento	Normalización por canal en entrenamiento	Escalado MinMax y ENN
Modelo	Dos bloques convolucionales y cabeza densa	Reglas difusas tipo 1 con 3 particiones
Decisión	Umbral 0.5	Umbral 0.77 para *Old*

3. Resultados

La comparación de rendimiento se muestra en la Fig. 1. La CNN compacta alcanzó el mejor resultado en todas las métricas, pero la diferencia con ExFuzzy fue pequeña en las medidas globales: 0.020 puntos en accuracy agrupado y 0.024 puntos en AUC. Esta cercanía es relevante porque el modelo difuso trabaja con una representación mucho más comprimida y, a cambio, conserva una lectura directa de sus decisiones.

La diferencia entre ambos modelos debe leerse junto con la naturaleza de sus entradas. La CNN recibe la matriz espectral completa, mientras que ExFuzzy opera sobre un número reducido de descriptores agregados. Que el margen en accuracy y AUC sea limitado sugiere que parte de la información relevante para la tarea queda capturada por variables compactas de conectividad, sin necesidad de utilizar una representación neuronal de mayor capacidad.

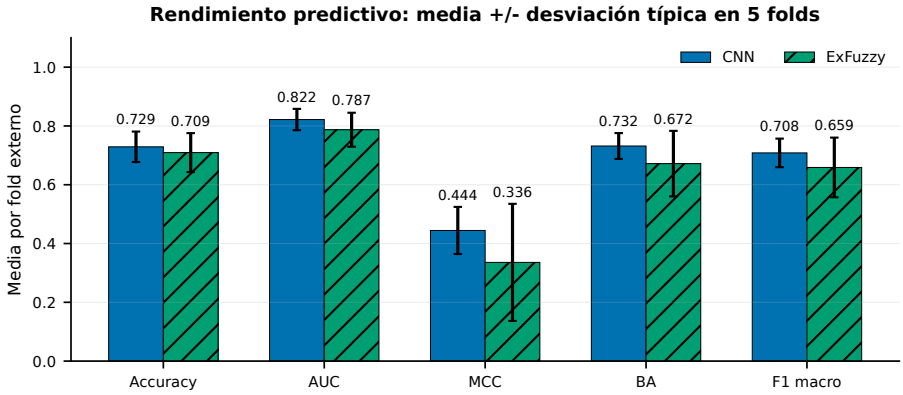


Figura 1. Comparación de rendimiento entre la CNN compacta y ExFuzzy. Las barras muestran media y desviación típica en los cinco pliegues externos. La CNN conserva ventaja predictiva, pero ExFuzzy se mantiene cerca con una estructura mucho más interpretable. BA denota balanced accuracy.

Más allá de las métricas agregadas, el interés del modelo difuso está en que permite inspeccionar el criterio de clasificación. La predicción no queda reducida a una puntuación final, sino que puede relacionarse con un conjunto pequeño de variables fisiológicamente interpretables. Esta propiedad facilita discutir si los patrones utilizados por el modelo son coherentes con el conocimiento disponible sobre conectividad y envejecimiento.

La diferencia cualitativa entre ambos enfoques aparece en la base final de reglas. La CNN devuelve una probabilidad, pero no una descripción directa de qué propiedades neurofisiológicas han pesado en la decisión. ExFuzzy, en cambio, utiliza cinco variables seleccionadas: *Theta_ Strength*, *Theta_ Centrality*,

Beta_ Entropy, *Gamma_ Strength* y *Gamma_ Entropy*. Las reglas asociadas a *Young* incluyen baja entropía beta, alta fuerza theta y una combinación de centralidad theta media con fuerza gamma media. Las reglas asociadas a *Old* incluyen baja fuerza theta y alta entropía gamma. La Fig. 2 resume estos antecedentes y muestra para cada regla los valores de *dominance score* (DS) y accuracy interna (ACC) proporcionados por ExFuzzy.

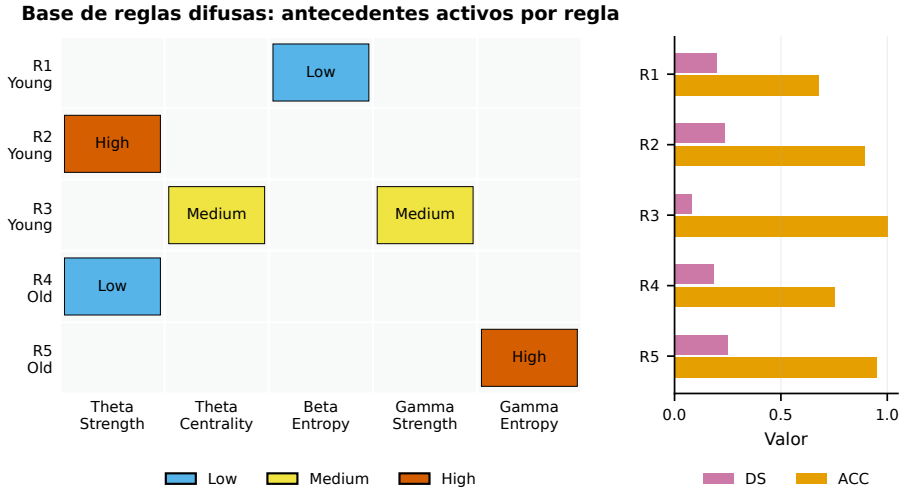


Figura 2. Visualización de la base de reglas final de ExFuzzy. La matriz izquierda indica qué término lingüístico participa en cada antecedente; el panel derecho muestra DS y ACC de cada regla. Esta representación permite inspeccionar de forma directa qué descriptores conducen a cada clase.

4. Discusión

Los resultados apuntan a una conclusión práctica: no todos los problemas de clasificación requieren necesariamente deep learning como primera respuesta. La CNN logra mejores métricas agregadas, algo esperable porque recibe una representación mucho más rica de la matriz espectral completa. El sistema difuso, en cambio, trabaja con una compresión drástica de la señal y aun así queda solo 0.020 puntos por debajo de la CNN en accuracy agrupado y 0.024 puntos por debajo en AUC. En un contexto donde solo importara exprimir la métrica, esa diferencia favorecería a la CNN. En un contexto donde la decisión debe ser examinada, discutida y relacionada con conocimiento del dominio, la cercanía del rendimiento cambia la lectura: un modelo interpretable puede ser una opción más adecuada aunque no alcance el máximo absoluto. Esta lectura es coherente con la defensa de modelos intrínsecamente interpretables en dominios de alto

impacto [11] y con la idea de que una explicación útil debe poder discutirse desde el conocimiento del usuario, no solo desde el funcionamiento interno del algoritmo [8].

La interpretación de las reglas exige cautela, pero también muestra por qué este tipo de framework es valioso. La presencia de fuerza theta, entropía beta y entropía gamma encaja con la idea general de que el envejecimiento sano modifica ritmos de reposo y patrones de conectividad [6]. Un modelo basado en reglas permite formular esa relación en términos que pueden revisarse: si baja la fuerza theta, si aumenta la entropía gamma, si una combinación de centralidad y fuerza entra en una zona media. Esa lectura no está disponible de forma inmediata en la CNN. Sin embargo, este estudio es predictivo y observacional; no permite atribuir causalidad ni establecer biomarcadores clínicos.

Desde el punto de vista metodológico, la comparación debe leerse como una demostración de trazabilidad, no como una evaluación definitiva de todas las arquitecturas posibles. Ni la CNN ni el sistema fuzzy fueron sometidos a una búsqueda extensa de hiperparámetros, por lo que las métricas reportadas describen estas configuraciones concretas y no el techo de rendimiento de cada familia metodológica. En trabajos futuros sería conveniente reforzar esta comparación con semillas múltiples, validación externa independiente, calibración del umbral dentro de cada fold y almacenamiento de predicciones por muestra, de modo que puedan analizarse las diferencias entre modelos con mayor precisión.

La literatura de sistemas difusos lingüísticos subraya que la interpretabilidad depende tanto del número de reglas como de la semántica de las variables utilizadas [12,4]. En este sentido, el modelo presentado cumple dos condiciones favorables: usa pocos descriptores y expresa las decisiones con términos lingüísticos simples. Esto no convierte automáticamente sus reglas en verdad neurobiológica, ni demuestra que constituyan la representación mínima, única o causalmente necesaria del fenómeno. La ventaja es otra: el modelo deja ver sus supuestos. Esa visibilidad permite discutirlos, contrastarlos con conocimiento experto y refutarlos si no se sostienen en nuevos datos. En campos como salud, finanzas o neurotecnología, esa posibilidad de auditoría puede ser tan importante como unas décimas de mejora en una métrica.

5. Conclusiones

Este estudio compara una CNN compacta y ExFuzzy en una tarea de clasificación de edad con EEG de reposo del conjunto LEMON. La CNN alcanza mejores métricas predictivas, pero ExFuzzy logra resultados muy próximos y expone su lógica mediante una base de reglas breve y visualizable. El mensaje principal es que deep learning no tiene por qué ser siempre la solución por defecto: cuando un framework interpretable sobre características biológicas descriptivas ofrece un rendimiento similar, la capacidad de explicar la decisión puede inclinar la balanza a su favor. La evidencia actual debe considerarse preliminar: las reglas son útiles para formular hipótesis sobre conectividad en bandas theta, beta y

gamma, pero requieren validación externa y análisis estadístico más amplio antes de interpretarse como marcadores neurofisiológicos estables.

Acknowledgments. Este trabajo ha sido financiado por la Agencia Estatal de Investigación (Gobierno de España, ayuda PID2022-136627NB-I00). Carmen Vidaurre cuenta con apoyo de las ayudas PID2020-118829RB-I00 y PID2024-161502OB-I00, y de IKERBASQUE.

Disclosure of Interests. Los autores declaran que no tienen intereses en conflicto relevantes para el contenido de este artículo.

Referencias

1. Arrieta, A.B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012
2. Babayan, A., Erbey, M., Kumral, D., Reinelt, J.D., Reiter, A.M.F., Roebbig, J., Schaare, H.L., Uhlig, M., Anwander, A., Bazin, P.L., et al.: A mind-brain-body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. Scientific Data **6**, 180308 (2019). https://doi.org/10.1038/sdata.2018.308
3. Fumanal-Idocin, J., Andreu-Perez, J.: Ex-Fuzzy: a library for symbolic explainable AI through fuzzy logic programming. Neurocomputing **599**, 128048 (2024). https://doi.org/10.1016/j.neucom.2024.128048
4. Gacto, M.J., Alcala, R., Herrera, F.: Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures. Information Sciences **181**(20), 4340–4360 (2011). https://doi.org/10.1016/j.ins.2011.02.021
5. Herrera, F.: Genetic fuzzy systems: taxonomy, current research trends and prospects. Evolutionary Intelligence **1**, 27–46 (2008). https://doi.org/10.1007/s12065-007-0001-5
6. Kang, J.H., Bae, J.H., Jeon, Y.J.: Age-related characteristics of resting-state electroencephalographic signals and the corresponding analytic approaches: a review. Bioengineering **11**(5), 418 (2024). https://doi.org/10.3390/bioengineering11050418
7. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. Journal of Neural Engineering **15**(5), 056013 (2018). https://doi.org/10.1088/1741-2552/aace8c
8. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
9. Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., Hallett, M.: Identifying true brain interaction from EEG data using the imaginary part of coherency. Clinical Neurophysiology **115**(10), 2292–2307 (2004). https://doi.org/10.1016/j.clinph.2004.04.029
10. Roy, Y., Banville, H., Albuquerque, I., Gramfort, A., Falk, T.H., Faubert, J.: Deep learning-based electroencephalography analysis: a systematic review. Journal of Neural Engineering **16**(5), 051001 (2019). https://doi.org/10.1088/1741-2552/ab260c
11. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x

12. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning—I. Information Sciences **8**(3), 199–249 (1975). https://doi.org/10.1016/0020-0255(75)90036-5