INTELIGENCIA ARTIFICIAL EXPLICABLE, IMPRESCINDIBLE EN EL ÁMBITO DE LA SALUD

La explicabilidad en inteligencia artificial (IA) se refiere a la capacidad de estos sistemas y modelos para proporcionar interpretaciones claras y comprensibles de cómo y por qué se han tomado determinadas decisiones o se han producido ciertos resultados. En otras palabras, es el proceso mediante el cual se hace transparente el funcionamiento interno de los algoritmos de IA, permitiendo a los usuarios, especialmente aquellos sin formación técnica avanzada, entender las bases de las predicciones y recomendaciones generadas por estos sistemas.

RAQUEL PODADERA RODRÍGUEZ,
Health of Marketing & Communications, Dedalus Iberia.

No podemos obviar que la IA y sus subcampos, como el Aprendizaje Automático (ML) y el Aprendizaje Profundo (DL), están revolucionando muchos aspectos del cuidado de la salud. Retos clínicos críticos relacionados con el diagnóstico, el pronóstico, la prevención, el descubrimiento de nuevos fármacos o los efectos de los tratamientos pueden mejorar considerablemente con su uso, además de ayudar en la personalización de evaluaciones clínicas y tratamientos para pacientes específicos, alineándose con la visión moderna de la medicina de precisión.

Grandes avances en la aplicación de la IA en salud

El apoyo al diagnóstico es una de las áreas más prometedoras de la IA en la salud. Mediante el análisis de grandes volúmenes de datos clínicos, los algoritmos pueden identificar patrones y anomalías complementarias al ojo humano. Por ejemplo, en el campo de la radiología, los sistemas de IA han demostrado ser capaces de detectar enfermedades como el cáncer de mama y o el de pulmón en imágenes de manera muy rápida y precisa. Esta capacidad para procesar y analizar datos a una velocidad y escala sin precedentes puede dar como resultado diagnósticos más tempranos y precisos.

Además del diagnóstico, el pronóstico es otra área donde la IA está teniendo un impacto significativo. Los modelos de ML pueden analizar datos históricos de pacientes para predecir la progresión de enfermedades y los posibles resultados de diferentes tratamientos. Esto permite a los médicos tomar decisiones más informadas sobre el manejo de enfermedades crónicas y la planificación del tratamiento a largo plazo. Por ejemplo, en la cardiología, los algoritmos de IA pueden predecir la probabilidad de eventos cardíacos futuros basándose en una variedad de factores, desde antecedentes médicos hasta datos de sensores portátiles.

El descubrimiento de fármacos también se ha beneficiado. Los algoritmos pueden analizar grandes bases de datos de compuestos químicos y datos biológicos para identificar nuevas combinaciones de fármacos y predecir su eficacia. Este enfoque ha acelerado el proceso de desarrollo de medicamentos, reduciendo el tiempo y el costo asociados con los ensayos clínicos tradicionales. Un ejemplo es el uso de IA por parte de empresas biotecnológicas para identificar posibles tratamientos para enfermedades como el COVID-19, donde la velocidad en el desarrollo de terapias es crucial.

La explicabilidad y su aportación a la ética

A pesar de estos avances, los modelos complejos, especialmente aquellos basados en Deep Learning (DL), presentan desafíos significativos debido a su naturaleza opaca. Estas "cajas negras" generan predicciones sin explicar los motivos detrás de sus resultados, lo cual es problemático en aplicaciones médicas donde las decisiones clínicas afectan directamente la salud humana. En este contexto, la explicabilidad en IA se ha convertido en una prioridad esencial para garantizar aplicaciones justas y éticas en la medicina. Es esencial que los Sistemas de Apoyo a la Decisión Clínica (CDSSs) impulsados por IA sean transparentes, para que los profesionales de la salud puedan entender el proceso y el fundamento de sus sugerencias. La opacidad en estos sistemas puede ocasionar problemas graves. Por ejemplo, una predicción equivocada sin una justificación clara puede conducir a decisiones médicas erróneas, poniendo en riesgo la seguridad del paciente.

Su importancia también se refleja en el marco legislativo y regulador. En este sentido, la Ley de Inteligencia Artificial de la UE (AI Act) establece que los sistemas de IA deben ser transparentes y comprensibles para los usuarios, mitigando la problemática de los modelos opacos. La AI Act identifica las aplicaciones de alto riesgo, como las médicas, donde la transparencia y la explicabilidad son críticas, asegurando que las decisiones clínicas basadas en IA sean comprensibles para los profesionales de la salud y promoviendo la confianza y la seguridad del paciente.

Además, la ley insiste en que los CDSSs deben proporcionar explicaciones claras y accesibles sobre cómo se generan las decisiones y recomendaciones. Reconociendo los riesgos asociados con la falta de explicabilidad, la AI Act establece medidas para que las predicciones y decisiones generadas por IA puedan ser verificadas y comprendidas, protegiendo así la seguridad de los usuarios. Esta ley complementa el Reglamento General de Protección de Datos (GDPR), que exige que las decisiones automatizadas sean explicables para proteger los derechos de los individuos, asegurando que las decisiones basadas en datos sean justificables y claras. Estas regulaciones impulsan la necesidad de desarrollar sistemas de IA que proporcionen claridad y justificación detrás de cada decisión, asegurando que los pacientes y los profesionales clínicos puedan confiar en estas tecnologías.

¿Qué aporta la explicabilidad?

En los últimos años, se han propuesto diversas metodologías para mejorar la explicabilidad de los modelos de IA en la salud. Entre estas, las Explicaciones Aditivas de Shapley (SHAP) han ganado popularidad por su capacidad para descomponer las contribuciones individuales de las características en las predicciones de modelos complejos. Este enfoque permite que los profesionales de la salud entiendan el impacto de cada variable en la predicción final, facilitando una toma de decisiones más informada.

La investigación en este campo no solo se enfoca en hacer que los modelos sean más transparentes, sino también en garantizar que estos sean justos y equitativos. Los modelos deben ser capaces de proporcionar explicaciones coherentes y comprensibles, independientemente del perfil del paciente, asegurando así que todos los individuos reciban un trato justo y adecuado.

La explicabilidad de los modelos de IA es crucial para los clínicos, ya que mejora la confianza en las decisiones asistidas por tecnología, optimiza la toma de decisiones y facilita la gestión de riesgos. Al hacer que los modelos de IA sean transparentes y comprensibles, se potencia su utilidad en la práctica clínica, permitiendo a los profesionales de la salud entender y justificar las recomendaciones proporcionadas por la IA. Esto no solo mejora la calidad de

la atención médica, sino que también asegura una integración ética y responsable de la tecnología en los procesos clínicos, alineándose con las normativas de protección de datos y las exigencias de transparencia establecidas por la Ley de Inteligencia Artificial de la UE.

Además, ayuda a identificar y corregir posibles sesgos en los modelos. Un modelo explicable puede revelar cómo ciertas características influyen en las predicciones, permitiendo a los desarrolladores ajustar los algoritmos para evitar decisiones sesgadas que podrían perjudicar a ciertos grupos de pacientes. Por ejemplo, se ha observado que algunos algoritmos de IA pueden estar sesgados en contra de minorías étnicas debido a la falta de representación en los datos de entrenamiento. La explicabilidad puede ayudar a detectar y mitigar estos sesgos, asegurando una aplicación más justa y equitativa de la tecnología.

Un ejemplo de cómo se está trabajando en desarrollar herramientas que aborden esta explicabilidad es el proyecto NEAR, participado por Dedalus, una solución modular que aborda la opacidad de los algoritmos ML/DL al convertir modelos complejos en explicables y transparentes sin una pérdida relevante de información. La metodología implementada en este proyecto es análoga al concepto de "destilación de conocimiento", que comprime modelos complejos en uno más simple y menos costoso computacionalmente. De esta forma, proporciona no solo una clasificación binaria, sino también un indicador de riesgo que representa la probabilidad de que ocurra una condición clínica, lo que es fácilmente interpretable como la suma de las contribuciones de las características individuales. Esto cumple con los requisitos fundamentales de explicabilidad, permitiendo a los profesionales clínicos identificar las características más importantes que constituyen el riesgo final y entender la relevancia de los valores faltantes en el puntaje final de riesgo.

Seguir avanzando en su desarrollo

A medida que la IA continúa avanzando, es fundamental que el ecosistema sanitario y los desarrolladores de tecnología trabajen juntos para desarrollar modelos que no solo sean precisos, sino también interpretables y transparentes, asegurando un futuro donde la IA mejore la atención al paciente de manera ética y responsable. La colaboración interdisciplinaria es clave en este esfuerzo. Los profesionales sanitarios pueden proporcionar la perspectiva clínica necesaria para guiar el desarrollo de algoritmos relevantes y útiles, mientras que los expertos en IA pueden diseñar modelos que sean más fáciles de entender y utilizar en el entorno clínico.

La IA tiene el potencial de transformar profundamente el cuidado de la salud, mejorando diagnósticos, pronósticos

y tratamientos. Sin embargo, para realizar este potencial de manera ética y efectiva, es crucial que estos sistemas sean explicables y transparentes. La explicabilidad no solo aumenta la confianza en estas tecnologías, sino que también asegura que las decisiones apoyadas de procesos automatizados sean justas y equitativas, protegiendo así los derechos y la seguridad de los pacientes.

Bibliografía

1. Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of Explainable AI Techniques in Healthcare. Sensors (Basel, Switzerland), 23. https://doi.org/10.3390/s23020634.
2. Mesinovic, M., Watkinson, P., & Zhu, T. (2023). Explainable AI for clinical risk prediction: a survey of concepts, methods, and modalities. ArXiv, abs/2308.08407. https://doi.org/10.48550/arXiv.2308.08407.
3. Kassem, K., Sperti, M., Cavallo, A., Vergani, A. M., Fassino, D., Moz, M., Liscio, A., Banali, R., Dahlweid, M., Benetti, L., Bruno, F., Gallone, G., De Filippo, O., Iannaccone, M., D'Ascenzo, F., De Ferrari, G. M., Morbiducci, U., Della Valle, E., & Deriu, M. A. (2024). An innovative artificial intelligence-based method to compress complex models into explainable, model-agnostic and reduced decision support systems with application. to healthcare (NEAR). Artificial Intelligence in Medicine, 151(102841), 102841. https://doi.org/10.1016/j.artmed.2024.102841
4. Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7
5. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery, 9(4). https://doi.org/10.1002/widm.1312
6. Sutton, R. T., Pincock, D., Baumgart, D. C., Sadowski, D. C., Fedorak, R. N., & Kroeker, K. I. (2020). An overview of clinical decision support systems: benefits, risks, and strategies for success. Npj Digital Medicine, 3(1), 1–10. https://doi