Combinando las fortalezas de los radiólogos y la IA para la detección del cáncer de mama: un análisis retrospectivo

Christian Leibig*, Moritz Brehmer*, Stefan Bunk, Danalyn Byng, Katja Pinker†,Lale Umutlu†

Resumen

Antecedentes Proponemos un enfoque de referencia de decisión para integrar la inteligencia artificial (IA) en la vía de detección del cáncer de mama, mediante el cual el algoritmo hace predicciones sobre la base de su cuantificación de la incertidumbre.
Las evaluaciones algorítmicas con alta certeza se realizan automáticamente, mientras que las evaluaciones con menor certeza se remiten al radiólogo. Este sistema de IA de dos partes puede clasificar exámenes de mamografía normales y proporcionar detección de cáncer post-hoc para mantener un alto grado de sensibilidad. Este estudio tuvo como objetivo evaluar el rendimiento de este sistema de IA en cuanto a sensibilidad y especificidad cuando se usa como un sistema independiente o dentro de un enfoque de referencia de decisión, en comparación con la decisión original del radiólogo.

Métodos Utilizamos un conjunto de datos retrospectivo que consta de 1 193197 estudios de mamografía digital de campo completo realizados entre el 1 de enero de 2007 y el 31 de diciembre de 2020, en ocho sitios de detección que participan en el programa nacional alemán de detección de cáncer de mama. Derivamos un conjunto de datos de prueba interna de seis sitios de detección (1670 cánceres detectados por detección y 19997 exámenes de mamografía normales), y un conjunto de datos de prueba externa de exámenes de detección de cáncer de mama (2793 cánceres detectados por detección y 80058 exámenes normales) de dos sitios de detección adicionales para evaluar el rendimiento de un algoritmo de IA sobre la sensibilidad y la especificidad cuando se usa como un sistema independiente o dentro de un enfoque de referencia de decisión, en comparación con la decisión original del radiólogo individual en el punto de lectura de la pantalla antes de la conferencia de consenso. Se evaluaron diferentes configuraciones del algoritmo de IA. Para tener en cuenta el enriquecimiento de los conjuntos de datos provocado por el sobremuestreo de los casos de cáncer, se aplicaron ponderaciones para reflejar la distribución real de los tipos de estudio en el programa de cribado. El desempeño del triaje se evaluó como la tasa de exámenes correctamente identificados como normales.

Se comparó la sensibilidad entre subgrupos clínicamente relevantes, sitios de detección y fabricantes de dispositivos entre la IA independiente, el radiólogo y la derivación de decisiones. Presentamos las curvas de características operativas del receptor (ROC) y el área bajo la ROC (AUROC) para evaluar el rendimiento del sistema de IA en todo su rango operativo. La comparación con los radiólogos y el análisis de subgrupos se basó en la sensibilidad y la especificidad en configuraciones clínicamente relevantes.

Hallazgos La configuración ejemplar del sistema de IA en modo independiente logró una sensibilidad del 84,2 % (95 % IC 82,4–85,8) y una especificidad del 89,5 % (89,0–89,9) en -datos de prueba, y una sensibilidad del 84,6 % (83,3–85,9) y una especificidad del 91,3 % (91,1–91,5) en datos de pruebas externas, pero fue menos preciso que el radiólogo promedio sin ayuda. Por el contrario, el enfoque de derivación de decisiones simuladas mejoró significativamente la sensibilidad del radiólogo en 2,6 puntos porcentuales y la especificidad en 1,0 puntos porcentuales, lo que corresponde a un rendimiento de triaje del 63,0 % en el conjunto de datos externo; el AUROC fue de 0,982 (IC del 95 %: 0,978–0,986) en el subconjunto de estudios evaluados por IA, superando el rendimiento del radiólogo. El enfoque de derivación de decisiones también produjo aumentos significativos en la sensibilidad para una serie de subgrupos clínicamente relevantes, incluidos los subgrupos de lesiones de pequeño tamaño y carcinomas invasivos. La sensibilidad del enfoque de derivación de decisiones fue consistente en los ocho sitios de detección incluidos y los tres fabricantes de dispositivos.

Interpretación El enfoque de derivación de decisiones aprovecha las fortalezas tanto del radiólogo como de la IA, demostrando mejoras en la sensibilidad y especificidad que superan las del radiólogo individual y del sistema de IA independiente. Este enfoque tiene el potencial de mejorar la precisión de la detección de los radiólogos, se adapta a los requisitos de la detección y podría permitir la reducción de la carga de trabajo antes de la conferencia de consenso, sin descartar el conocimiento generalizado de los radiólogos.

Financiación Vara.

Lancet Digit Salud 2022;
4: e507-19

Ver página de comentarios e478

*Primeros autores conjuntos

†Últimos autores conjuntos

Vara, Berlín, Alemania
(C Leibig PhD, M Brehmer MD,
S Litera MSc, D Byng MSc);
Departamento de Diagnóstico y
Radiología Intervencionista y
Neurorradiología, Universidad
Hospital Essen, Essen, Alemania
(M Brehmer, L Umutlu MD);
Departamento de Radiología,
servicio de imagen mamaria,
Monumento a Sloan Kettering
Centro de Cáncer, Nueva York, NY,
EE. UU. (K Pinker MD); Departamento de
Imágenes Biomédicas y
División de Terapia Guiada por Imágenes
Molecular y de Género
Imágenes, Universidad de Medicina de
Viena, Viena, Austria
(K Pinker)

Correspondencia a: Dr.
Christian Leibig, Vara, 13355 Berlín,
Alemania
christian.leibig@vara.ai

Introducción

El aumento de la popularidad de las redes neuronales profundas (DNN) en imágenes médicas, provocado por los avances en inteligencia artificial (IA) para el reconocimiento de imágenes y la mayor disponibilidad de datos de mamografía digital, han suscitado interés en nuevos modelos basados en características de imágenes cuantitativas para mejorar interpretación de mamografías.1

Los estudios recientemente publicados sobre la detección y clasificación de lesiones basadas en DNN sobre la base de datos de mamografía digital han demostrado que dichos sistemas tienen un rendimiento de diagnóstico comparable al de los radiólogos y son prometedores como sistemas de apoyo a la toma de decisiones,2–8, pero la evidencia actual es insuficiente para juzgar precisión dentro de los programas de detección del cáncer de mama.9

Investigación en contexto

Evidencia antes de este estudio

Hicimos una búsqueda bibliográfica en PubMed de artículos en inglés desde el inicio hasta el 1 de septiembre de 2021, para identificar estudios que usaron "aprendizaje automático", "aprendizaje profundo", "mamografía" y "detección de cáncer de mama" para clasificar exámenes normales o detectar lesiones sospechosas en mamografías de detección, o que utilizaron una técnica de referencia de decisión para mejorar los parámetros de eficacia de la detección a través de un enfoque humano en el circuito, en poblaciones típicas de detección de cáncer de mama. Identificamos algoritmos de inteligencia artificial (IA) para la clasificación de los exámenes de detección del cáncer de mama que se centraron en tomar todas las decisiones, incluidos los casos no concluyentes, sin considerar los efectos posteriores. En nuestra búsqueda, identificamos algoritmos de IA que usaban sistemas de clasificación enfocados en proporcionar solo predicciones negativas, porque se ha postulado que dichos sistemas podrían aumentar potencialmente la sensibilidad de la detección del cáncer al permitir que los radiólogos tengan más tiempo para revisar los casos graves. Sin embargo, tales sistemas, aunque reducen en gran medida la carga de trabajo del radiólogo, también podrían disminuir la sensibilidad para la detección del cáncer. Otros enfoques de IA se centran en la interpretación independiente de las mamografías de detección para automatizar por completo la detección de lesiones sospechosas. Sin embargo, debido a la baja prevalencia de cánceres en las mamografías de detección, tales sistemas completamente automatizados generarán una gran cantidad de falsos positivos, lo que requerirá más recursos de atención médica para clasificar de manera segura los estudios normales. Por el contrario, tales enfoques podrían afectar la sensibilidad del radiólogo al hacer predicciones positivas excesivas sobre hallazgos benignos y distraer la atención de los cánceres reales. Un estudio publicado exploró la combinación de preselección y selección mejorada (suplementaria) para estudios negativos con puntajes de IA altos; sin embargo, este estudio no consideró explícitamente el efecto de la normalidad triaje y detección de cáncer en el momento de la mamografía en la sensibilidad y especificidad del radiólogo.

Valor agregado de este estudio

Este estudio tuvo como objetivo proponer una solución hacia la adopción clínica segura de sistemas de IA en la detección del cáncer de mama. Es decir, este estudio apoyó la adopción de un enfoque colaborativo de IA-radiólogo que combina el triaje y la detección de cáncer con alta precisión, y renuncia a un enfoque de IA independiente que tiene como objetivo reemplazar al radiólogo, pero con el riesgo de degradar la sensibilidad. En particular, proponemos un enfoque de derivación de decisiones que aprovecha las fortalezas tanto del radiólogo como del algoritmo de IA. Si el algoritmo funciona con mayor precisión en un subconjunto de estudios y el radiólogo es mejor en el otro, cada uno puede realizar predicciones en las que se destacan. Este sistema de dos partes incorpora la clasificación de los exámenes y una red de seguridad para predecir los exámenes positivos para el cáncer a fin de mantener un alto grado de sensibilidad para la detección del cáncer, con la red de seguridad sirviendo como soporte de decisiones post-hoc para el radiólogo. Este enfoque mejora la precisión de detección de los radiólogos, se adapta a los requisitos de detección y permite reducir la carga de trabajo de los radiólogos sin descartar su supervisión final.

Implicaciones de toda la evidencia disponible Los resultados de este estudio podrían mejorar la implementación segura de los algoritmos de IA, lo que conduciría a mejores parámetros de eficacia en los programas de detección a nivel nacional y reduciría la carga de trabajo de los radiólogos. Mostramos que las configuraciones realistas de nuestro algoritmo de IA dentro de un enfoque de referencia de decisiones mejoraron las métricas de detección no solo en promedio sino también en subgrupos clínicamente relevantes. Además, este enfoque podría generalizarse a datos de sitios de detección no vistos.

El trabajo anterior ha demostrado el potencial de combinar las fortalezas de los radiólogos y los modelos de aprendizaje automático utilizando métodos de aprendizaje en conjunto, consolidando las predicciones de los radiólogos y los modelos.7,8 Sin embargo, un inconveniente importante de este enfoque es la necesidad de que el radiólogo evalúe todos los estudios. , por lo que la IA no alivia la carga de trabajo del radiólogo. Otro trabajo evaluó un enfoque de clasificación impulsado por IA para la detección, mediante el cual los exámenes con una alta probabilidad de estar libres de cáncer se clasifican y los exámenes restantes se derivan al radiólogo.10–15 Estos estudios, sin embargo, mostraron que las reducciones considerables de los exámenes de detección de la carga de trabajo del radiólogo podría derivar en una reducción inaceptable de la sensibilidad.

Una solución comercial incorporó el triaje normal en un paso seguido de la identificación de mujeres en riesgo de falsos negativos que tenían una doble lectura negativa, pero que podrían beneficiarse de una evaluación mejorada con imágenes complementarias con MRI o ultrasonido.10 Aunque este enfoque mejoró indirectamente la sensibilidad de la detección del cáncer, solo se centró en predicciones sobre el riesgo futuro de cáncer de intervalo o la siguiente ronda de cánceres detectados por exámenes de detección que ambos lectores pasaron por alto, y no se centró en las predicciones sobre los exámenes positivos para el cáncer visibles en la mamografía en el examen de detección en sí, lo que podría pasar por alto por uno de dos lectores en un Configuración de doble lector. Por lo tanto, hasta la fecha, ningún estudio ha explorado la IA que combine la clasificación normal y la detección del cáncer en el punto de lectura de la pantalla de mamografía por parte de radiólogos individuales, antes de la conferencia de consenso, y el efecto de dicho enfoque en la sensibilidad y especificidad del radiólogo. Comprender cómo un sistema de IA de este tipo podría afectar las métricas de detección del radiólogo requiere una ilustración de cómo estos dos sistemas deben trabajar juntos para lograr una mejora conjunta de la sensibilidad y la especificidad, dado que el aumento de la sensibilidad generalmente se produce a expensas de la reducción de la esp viceversa

En busca de un sistema basado en IA que pueda ser utilizado por lectores individuales antes de las reuniones de consenso o arbitraje, que al mismo tiempo mejore la sensibilidad del lector y mantenga o incluso mejore la especificidad con seguridad

clasificando estudios normales, proponemos un sistema de IA que utiliza un enfoque de referencia de decisión. Este enfoque de derivación de decisiones realiza evaluaciones algorítmicas muy confiables automáticamente, mientras que las evaluaciones menos confiables se derivan al radiólogo. Este sistema de dos partes incorpora el triaje de exámenes normales y al mismo tiempo introduce una red de seguridad para mantener un alto grado de sensibilidad al realizar predicciones en exámenes positivos para cáncer. Este sistema está diseñado para ser utilizado por el radiólogo individual que lee la mamografía de detección, antes de la revisión por consenso y, por lo tanto, se garantiza la evaluación de su rendimiento en los cánceres detectados por detección y los exámenes de mamografía normales comprobados de seguimiento.

Para mejorar el rendimiento del diagnóstico, primero tenemos que demostrar que las predicciones fiables16 del modelo, que permitirían evaluar estos estudios de forma totalmente automática sin derivarlos al radiólogo, superarían a las del lector humano. Aquí, describimos el desarrollo y la evaluación de dicho algoritmo de clasificación de cáncer basado en DNN, utilizando un conjunto de datos de 1193197 estudios de detección derivados de un programa nacional de detección de cáncer de mama. Presumimos que el modelo sería lo suficientemente sensible y específico para clasificar de forma independiente los casos normales y reconocer los casos sospechosos. Además, nuestro objetivo era demostrar la mejora de la precisión diagnóstica de detección (sensibilidad y especificidad) del radiólogo cuando se utiliza el enfoque de derivación de decisiones, con posibilidad de generalización en diferentes sitios de detección y fabricantes de dispositivos. El desempeño del enfoque de referencia de decisión se contrastó aún más con el desempeño del algoritmo de IA en modo independiente.

Métodos

Diseño del estudio

En este estudio de análisis retrospectivo, se comparó el rendimiento de detección de un solo radiólogo sin ayuda basado en sus decisiones clínicas originales en el programa de detección (figura 1A) con el de un sistema de IA independiente (figura 1B) y un enfoque de derivación de decisiones (figura 1C) que combina el triaje normal y la detección del cáncer a través de un sistema de alerta de red de seguridad.

Las decisiones del radiólogo original fueron las registradas durante la práctica clínica sin apoyo de IA en el punto de lectura de pantalla antes de la conferencia de consenso o arbitraje. Por lo tanto, los análisis en este estudio se limitaron a los cánceres detectados por exámenes de detección y a los exámenes de mamografía normales comprobados durante el seguimiento.

Simulamos un escenario de cribado (figura 1C) en el que, en un primer paso, el sistema de IA clasificaba si un estudio era normal o sospechoso de cáncer y proporcionaba al mismo tiempo una indicación de su confianza en su clasificación, sobre la base de dos umbrales.17 Tanto los estudios sospechosos como los estudios para los cuales el algoritmo no era seguro y requerían interpretación humana fueron remitidos al radiólogo sin indicación del

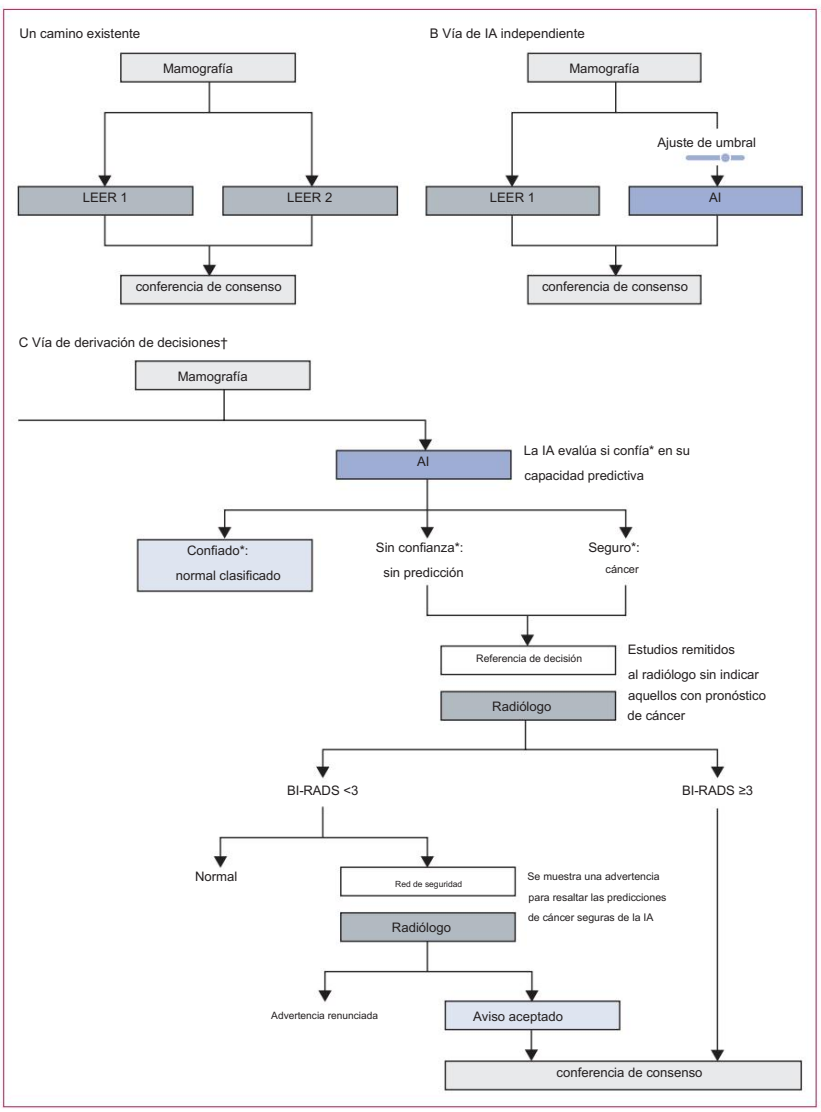

Figura 1: Comparación entre la vía de referencia de decisión y la vía de IA independiente en entornos de detección de doble lector Se presentan diferentes vías de detección posibles. (A) La vía de detección existente, en la que los estudios de mamografía son revisados de forma independiente por dos lectores y los hallazgos discordantes se resuelven durante el consenso. (B) La vía de IA independiente, la vía de implementación más comúnmente propuesta para los sistemas de IA. Independiente se define por la toma de todas las decisiones de un radiólogo, a veces también denominado lectura independiente. (C) La vía de decisión-remisión, que es el enfoque de esta evaluación. Todos los estudios de mamografía son leídos primero por el sistema de IA y se producen las predicciones. IA = inteligencia artificial. *El modelo presenta una puntuación entre 0·0 y 1·0 que indica la malignidad de un estudio. Las puntuaciones inferiores al umbral para las predicciones negativas (clasificadas como normales) o superiores al umbral para las predicciones positivas (red de seguridad) se consideraron seguras. Todas las demás puntuaciones entre los dos umbrales no se consideraron fiables y los estudios correspondientes se remitieron al radiólogo. †Enfoque de referencia de decisión cuando lo utiliza un solo lector en un entorno de doble lector.

Clasificaciones del sistema de IA. Además, evaluamos una red de seguridad, que fue activada por estudios que el sistema de inteligencia artificial consideró sospechosos de cáncer.

Supuestos de simulación

Debido a la naturaleza retrospectiva de este estudio, evaluamos la sensibilidad y la especificidad para la detección del cáncer a partir del escenario en el que se

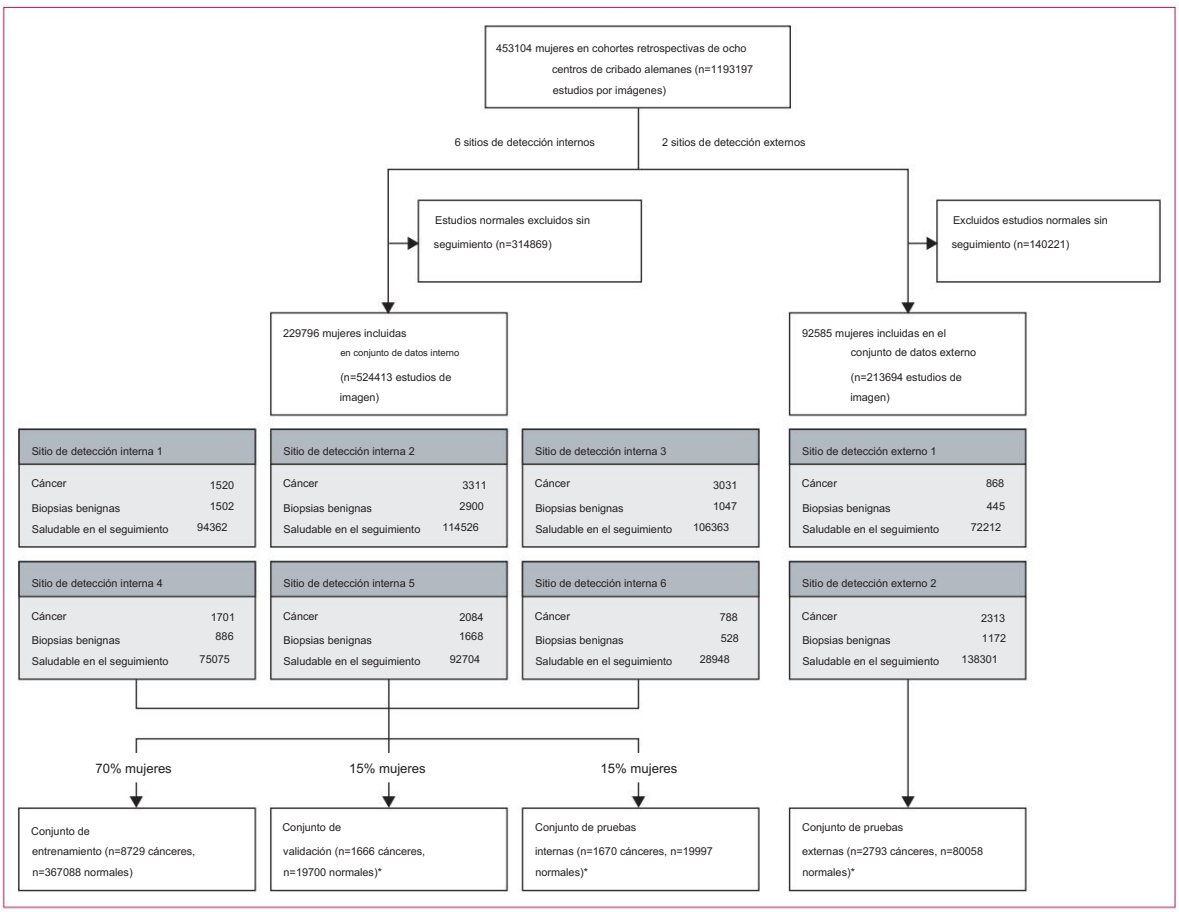

Figura 2: Particiones del conjunto de datos En el apéndice (pág. 6) se encuentra disponible más información sobre los criterios de inclusión del estudio, el programa nacional alemán de detección del cáncer de mama y la técnica de ponderación de la muestra. *Exámenes de mamografía normales de submuestra, un estudio por mujer.

El radiólogo acepta las clasificaciones del modelo de IA de los estudios normales y de red de seguridad clasificados, mientras que las clasificaciones de los estudios restantes se basaron en las decisiones del radiólogo. Esto es equivalente a modelar las predicciones confiables de IA como completamente automáticas (es decir, no es necesario mostrar predicciones de IA al radiólogo) y, por lo tanto, nos permitió evitar formular hipótesis y dar cuenta de las interacciones humano-IA.

Fuentes de datos

Este estudio fue revisado por abogados especializados en privacidad de datos para garantizar el cumplimiento del Reglamento general de protección de datos de la UE. La aprobación ética y la necesidad de obtener el consentimiento informado no se aplicaron a este estudio en virtud de la legislación regional y nacional debido a la naturaleza retrospectiva y totalmente anonimizada de los estudios de mamografía y los datos de los pacientes.

Ver en línea para el apéndice

Utilizamos un conjunto de datos retrospectivo que consta de 1193197 estudios de mamografía digital de campo completo realizados entre el 1 de enero de 2007 y el 31 de diciembre de 2020, de 453104 mujeres, datos que se recuperaron de ocho sitios de detección alemanes.

Derivamos un conjunto de datos de prueba interna a partir de seis evaluaciones y un conjunto de datos de pruebas externas de exámenes de detección de cáncer de mama de dos sitios de detección adicionales (figura 2). Todos los estudios de mamografía se realizaron con fines de detección en mujeres asintomáticas que se presentaron al programa nacional de detección de mamas; no se utilizaron imágenes de diagnóstico o de recuerdo. Los estudios sospechosos que entraron en la conferencia de consenso, incluidos los que se recordaron y se sometieron a biopsia, se sobremuestrearon durante la recopilación de datos, pero este enriquecimiento se abordó durante la evaluación del modelo con un enfoque de ponderación descrito en la sección de análisis estadístico. Todos los cánceres en el conjunto de datos se detectaron mediante detección; no se incluyeron los cánceres perdidos o diagnosticados en el intervalo entre las rondas de detección. Los exámenes de mamografía normales se derivaron de mujeres con exámenes de seguimiento dentro de un mínimo de 24 meses, que no se recordaron (BI-RADS 1 o 2) o, en el caso de un hallazgo, el estudio de seguimiento debe haber sido considerado negativo ya sea por lectura doble, conferencia de consenso o recuerdo negativo (apéndice p 2). Todos los estudios comprendían cuatro vistas estándar, craneocaudal bilateral y oblicua mediolateral. En cuanto a los fabricantes de dispositivos, el 43,1% de los estudios de mamografía se obtuvieron utilizando

un dispositivo Siemens, un 36,2% un dispositivo Hologic y un 8,4% un dispositivo Fuji. El 12,3% restante de los dispositivos de mamografía se obtuvieron utilizando dispositivos fabricados por otros fabricantes; estos se incluyeron en el conjunto de datos de entrenamiento, pero se excluyeron de las evaluaciones posteriores. Las mujeres tenían entre 50 y 70 años de edad en el momento de la selección; a más del 80 % de las mujeres se les asignaron categorías de densidad mamaria del American College of Radiology (ACR) B o C (apéndice p 2).

Los datos obtenidos en seis sitios de detección se usaron como un conjunto de datos interno, divididos aleatoriamente por ID de paciente en conjuntos de datos de capacitación, validación y prueba, siguiendo la práctica estándar para desarrollar y evaluar modelos de aprendizaje automático.18 Cada división fue mutuamente excluyente ; por lo tanto, las mujeres cuyos datos se usaron para el entrenamiento del modelo (70 %) y la validación (15 %) no se incluyeron en el conjunto de datos de prueba (15 %). Los conjuntos de datos de entrenamiento y validación se utilizaron para desarrollar el sistema de IA.
Los datos de validación se utilizaron para configurar los umbrales de referencia de decisión (apéndice p 3).

Utilizamos dos conjuntos de datos para evaluar el rendimiento del algoritmo, el conjunto de datos de prueba interna y el conjunto de datos de prueba externa. El conjunto de datos de la prueba interna constituyó una muestra independiente de mujeres que no se incluyeron en los conjuntos de datos de capacitación o validación, aunque procedían de los mismos seis sitios de detección utilizados para desarrollar el algoritmo. Para verificar que el rendimiento logrado del algoritmo no fue causado por el aprendizaje abreviado19 de señales específicas de esos seis sitios de detección, sino que se generalizó a diferentes sitios de detección, complementamos esta evaluación con una evaluación de datos fuera de distribución de dos sitios de detección adicionales. nunca antes visto por el sistema de IA. Para dar cuenta del enriquecimiento de cada conjunto de datos causado por casos de cáncer sobremuestreados, utilizamos una técnica de ponderación20,21 para garantizar que los conjuntos de datos de prueba reflejaran una población real de detección (apéndice p 6).

Desarrollo del algoritmo de IA El algoritmo de IA clasifica el cáncer a nivel de estudio. Solo se necesitan etiquetas y predicciones a nivel de estudio para evaluar la viabilidad de la derivación de decisiones. Si el enfoque de referencia de decisiones puede mejorar las métricas de detección depende de si el modelo puede hacer mejores predicciones que los radiólogos en un subconjunto de estudios. Presentamos un modelo basado en una red neuronal convolucional profunda, entrenado con imágenes de mamografía utilizando etiquetas en diferentes escalas (parche, imagen y estudio) solo con fines de entrenamiento. Esas etiquetas se derivaron de anotaciones de hallazgos radiológicos e información de biopsia asociada.

Los hallazgos de imágenes que fueron confirmados por biopsia fueron anotados por radiólogos certificados por la junta. Estos comprendían hallazgos radiológicos que inicialmente se clasificaron como sospechosos (BI-RADS 4 o 5, es decir, sospechosos o muy sospechosos de malignidad) y que luego se recomendaron para biopsia en la evaluación, y hallazgos radiológicos que inicialmente se clasificaron como BI-RADS 2 o 3 y eso después se sometieron a biopsia según la preferencia de los pacientes. Los radiólogos utilizaron un visor de radiología basado en la web dedicado, lo que les permitió el acceso simultáneo a los informes de histopatología y radiología. Los datos relacionados con el estándar de referencia histopatológico se extrajeron de los informes almacenados en el software de detección oficial del programa de detección alemán. Los informes se estandarizaron de acuerdo con la cuarta edición de las Directrices europeas para la garantía de calidad en la detección y el diagnóstico del cáncer de mama.22 Los estudios se etiquetaron como positivos en función de la confirmación histopatológica.23,24 Los radiólogos segmentaron cada región sospechosa en las imágenes respectivas con un polígono. . La arquitectura del modelo y la capacitación se describen en el apéndice (pág. 8).

Evaluación del algoritmo de IA El conjunto de datos de la prueba interna contenía 1670 casos de cáncer detectados por exámenes de detección y confirmados por biopsia y 19997 exámenes de mamografía normales comprobados durante el seguimiento, mientras que el conjunto de datos de pruebas externas contenía 2793 casos de cáncer detectados por exámenes de detección y 80058 casos de cáncer detectados por exámenes de seguimiento. exámenes de mamografía normales. Estos conjuntos de datos se utilizaron para evaluar el rendimiento del enfoque de IA independiente (figura 1B) y el enfoque de referencia de decisiones (figura 1C). La IA independiente se refiere a que la IA se hace cargo de todas las decisiones del radiólogo (es decir, no se remiten decisiones). El enfoque de derivación de decisiones combina predicciones algorítmicas confiables que no se derivan al radiólogo con la derivación de estudios menos confiables al radiólogo, con la hipótesis de que esta estrategia mantiene o mejora las métricas de detección clave, como las características de IA de una red de seguridad y El triaje normal logra una mejora global complementaria de la sensibilidad y especificidad del radiólogo. El enfoque de referencia de decisión pasa naturalmente al modo independiente cuando todas las predicciones algorítmicas se consideran confiables. La sensibilidad del radiólogo se refiere a la cantidad de cánceres detectados por la pantalla que encontró el radiólogo individual, dividido por la cantidad total de cánceres detectados por la pantalla en el conjunto de datos, que es la sensibilidad como porcentaje de la sensibilidad de lectura doble, con dos lectores que encuentran el 100 %. de los cánceres detectados por la pantalla.

Configuración de los enfoques de referencia de decisiones e IA independiente
El enfoque de referencia de decisiones se configuró de la siguiente manera: se establecieron umbrales inferiores (apéndice p 4) para predicciones negativas seguras (triaging normal) y umbrales superiores para predicciones positivas (red de seguridad) de modo que se la sensibilidad y la especificidad se lograron en el conjunto de datos de validación (figura 2). Dados dos umbrales, calculamos la sensibilidad y la especificidad generales del sistema combinado sobre la base de las evaluaciones de IA en estudios confiables y evaluaciones de radiólogos en estudios no confiables. Las sensibilidades y especificidades resultantes en los datos de validación se usaron para elegir la compensación deseada de sensibilidad y especificidad. Una configuración clínicamente significativa maximiza la sensibilidad del radiólogo sin

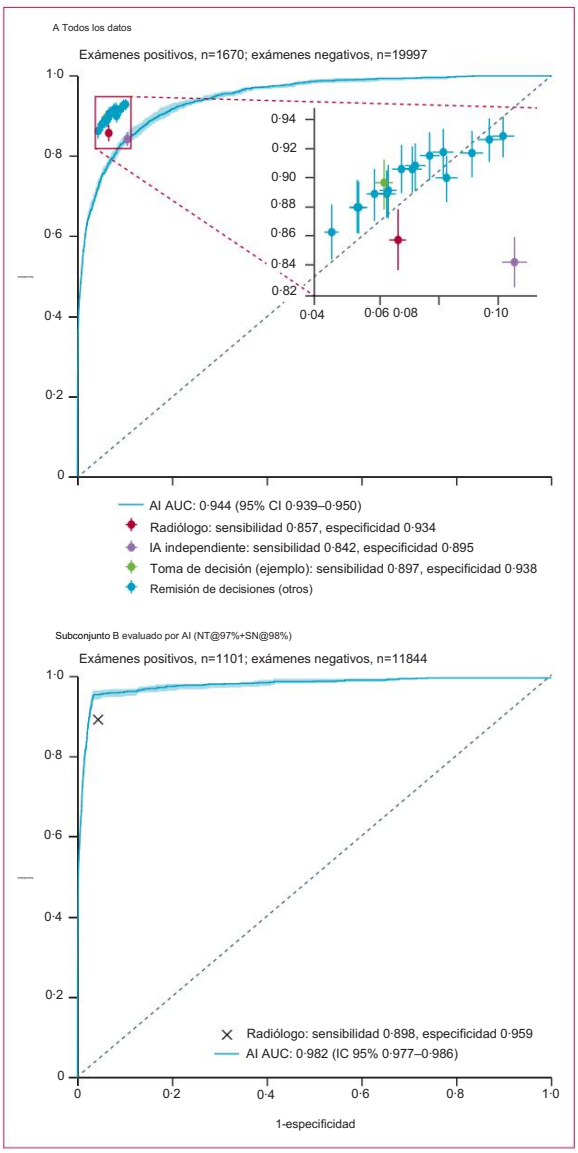

Figura 3: Comparación del rendimiento de los enfoques independientes y de referencia de decisión basados en el conjunto de datos de pruebas internas Se presenta la precisión diagnóstica de detección general para radiólogos, la IA independiente y la referencia de decisión. Se proporcionan sensibilidad y especificidad para radiólogos (rojo), IA independiente (púrpura) y referencia de decisión (verde para la configuración ejemplar NT@97%+SN@98% y azul para configuraciones alternativas). Además, presentamos curvas ROC y AUROC para evaluar el rendimiento del sistema de IA en todo su rango operativo en el conjunto de datos de prueba interna (n=21667; A) y en el subconjunto de datos para el que puede producir sus predicciones más confiables para la configuración ejemplar NT@97%+SN@98% (B). Las barras de error indican IC del 95 %. El enfoque de derivación de decisiones superó al radiólogo independiente en sensibilidad y especificidad, o en ambas, dependiendo de la configuración (A) al superar al radiólogo en todo el conjunto de predicciones confiables (B). Los valores de sensibilidad y especificidad resultantes para todos los estudios fueron similares o mayores que los del radiólogo solo, mientras que entre el 42,1 y el 71,1 % de los estudios pudieron clasificarse de manera segura. IA = inteligencia artificial. AUC=área bajo la curva. AUROC=área bajo la característica operativa del receptor. NT = triaje normal. ROC = característica de funcionamiento del receptor. SN=red de seguridad.

especificidad decreciente. En el conjunto de datos de validación, una sensibilidad algorítmica del 97 % y una especificidad del 98 % fue la mejor compensación lograda (apéndice p 3). La configuración que logró esta sensibilidad y especificidad se usa de manera ejemplar para presentar los resultados principales aquí, mientras que otras configuraciones se muestran en la tabla. Para cuantificar la reducción de la carga de trabajo, el rendimiento del triaje se calculó como la tasa de estudios etiquetados correctamente como normales (es decir, la fracción que podría automatizarse).

La IA independiente se configuró mediante el establecimiento de un umbral único (apéndice p 3) de modo que se lograra la sensibilidad del radiólogo del 86 % en el conjunto de datos de validación (figura 2).

Análisis estadístico

Las curvas para la característica operativa del receptor (ROC) y las áreas bajo la ROC (AUROC) se utilizaron como métricas para evaluar el rendimiento de la IA independiente en todo su rango operativo. Para determinados puntos operativos de IA independiente, el radiólogo y el enfoque de referencia de decisión, se calcularon estimaciones de sensibilidad y especificidad.

Para las estimaciones puntuales que involucran una decisión del radiólogo, se promediaron las dos decisiones independientes por estudio. Para las estimaciones de error y las pruebas de hipótesis se utilizaron métodos de remuestreo. Para todas las métricas estimadas, los IC del 95 % se determinaron sobre la base de 1000 muestras de arranque.25 La variabilidad del juicio humano influye en el radiólogo y en las métricas de referencia de decisiones y se tuvo en cuenta mediante un procedimiento de muestreo de dos pasos de la siguiente manera: para cada estudio mamográfico , la evaluación de un radiólogo se tomó como muestra de dos lectores independientes y anónimos; y todo el conjunto de datos se volvió a muestrear con reemplazo.

Para comprender si la adición de IA tuvo un efecto uniforme sobre la sensibilidad en los subgrupos clínicamente relevantes, calculamos los valores de sensibilidad específicos de los subgrupos en los conjuntos de datos de pruebas internas y externas según diferentes niveles de puntuación de biopsia, densidad mamaria ACR, tamaño de la lesión y hallazgos radiológicos de acuerdo con BI RADS.23,24,26 La capacidad de generalización se evaluó de manera similar al comparar los resultados en los conjuntos de datos de prueba y validación interna, y al presentar la sensibilidad estratificada entre los sitios de detección y los fabricantes de dispositivos, y la especificidad estratificada entre los fabricantes de dispositivos.

Se evaluó la significación estadística de las diferencias en la sensibilidad y la especificidad de la IA independiente frente al radiólogo y la derivación de decisiones frente al radiólogo mediante una prueba de permutación.25 Para cada uno de los 10000 ensayos, al igual que para los IC, una de las dos decisiones del radiólogo se muestreó de forma independiente para cada estudio mamográfico, y cada decisión de referencia de decisión se permutó aleatoriamente con la decisión del radiólogo. Se calculó un valor de p de dos lados al comparar la diferencia observada con los cuantiles de la distribución nula .

se describe en el apéndice (pág. 6).20,21 Los análisis se realizaron utilizando la pila de computación científica Python versión 3.8.10.

Papel de la fuente de
financiación El financiador del estudio participó en la recopilación, gestión y análisis de los datos utilizados para desarrollar el algoritmo de IA y en la preparación y revisión del manuscrito. Los autores no empleados por el financiador tenían el control de los datos y la información enviada para publicación en todo momento y tomaban la decisión final de enviar el manuscrito para publicación.

Resultados

El rendimiento del sistema de IA independiente se contrasta con el rendimiento del radiólogo (figura 3A, 4A).
El rendimiento del sistema de IA independiente en todas las configuraciones posibles se muestra mediante la curva ROC correspondiente, alcanzando un AUROC de 0,944 (IC del 95 %: 0,939–0,950) en el conjunto de datos de prueba interna, y 0,951 (0·947–0·955) en el conjunto de datos de prueba externa. En el conjunto de datos de la prueba interna, el radiólogo logró una sensibilidad del 85,7 % (95 % IC 83,6–87,9) y una especificidad del 93,4 % (95% IC 93,1–93,7), en comparación con una sensibilidad del 84,2 % (82,4–85,8) y una especificidad del 89,5 % (89,0–89,9) para el punto operativo del sistema de IA independiente que mantuvo la sensibilidad del radiólogo en el conjunto de datos de validación (figura 3A, 4A; tabla). En el conjunto de datos de la prueba externa, el rendimiento del radiólogo en comparación con la IA independiente fue del 87,2 % (85,6–88,7) frente al 84,6 % (83,3–85,9) en sensibilidad y del 93,4 % (93·2–93·6) versus 91·3% (91·1–91·5) en especificidad. La sensibilidad y especificidad del sistema de IA independiente fue significativamente más baja que la del radiólogo sin ayuda en ambos conjuntos de datos de prueba (p = 0,0019 para la sensibilidad de los datos de la prueba externa y p <0,0001 para la especificidad de los datos de la prueba interna y externa), pero la sensibilidad no fue significativamente diferente en el conjunto de datos de la prueba interna (p=0·17).

El rendimiento del enfoque de decisión-remisión se traza con puntos de mira (figuras 3A, 4A). Utilizando la configuración ejemplar, el enfoque de decisión-referencia logró una sensibilidad del 89,7 % (87,9–91,3) y una especificidad del 93,8 % (93,6–94,1), lo que representó un 4,0 mejora de un punto porcentual en la sensibilidad y una mejora de 0,5 puntos porcentuales en la especificidad en comparación con el radiólogo sin ayuda en el conjunto de datos de la prueba interna (tabla).
Este hallazgo correspondió a un rendimiento de triaje del 60,7 % y una mejora estadísticamente significativa tanto de la sensibilidad como de la especificidad (sensibilidad p<0,0001; especificidad p=0,0002). En el conjunto de datos de prueba externa, el enfoque de referencia de decisión logró una mejora significativa tanto en la sensibilidad (2,6 puntos porcentuales) como en la especificidad (1,0 punto porcentual; p<0,0001 para ambos), lo que corresponde a un desempeño de triaje al 63·0%.

Se muestran otras configuraciones posibles (figuras 3A, 4A; tabla). El enfoque de decisión-derivación superó al radiólogo sin ayuda tanto en sensibilidad como en especificidad.
Las configuraciones para las cuales la referencia de decisión tuvo un efecto diferente en la sensibilidad y la especificidad también son

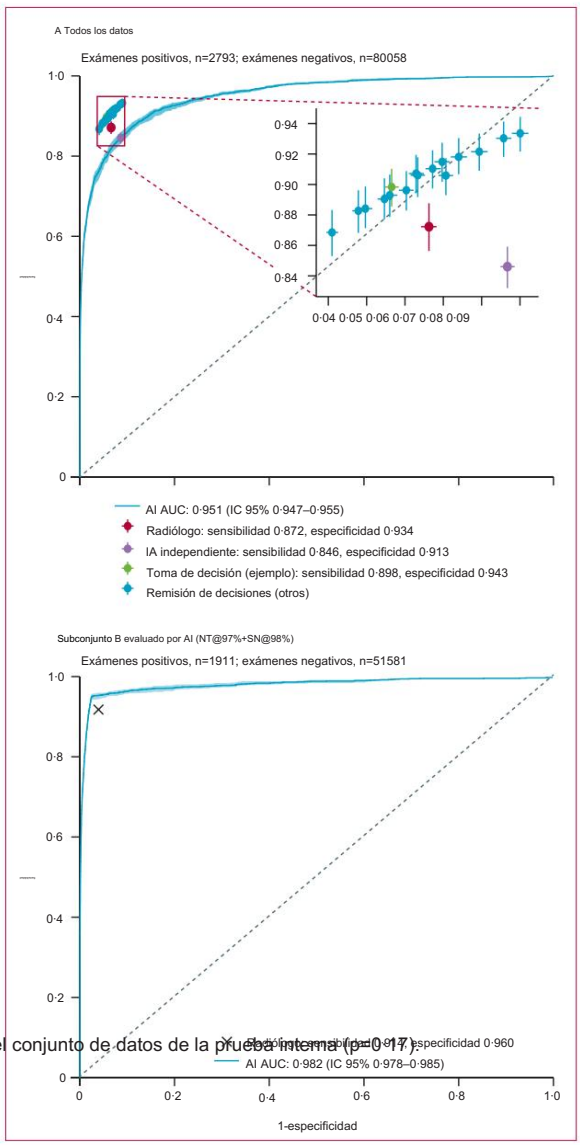

Figura 4: Comparación del rendimiento de los enfoques independientes y de referencia de decisión basados en el conjunto de datos de pruebas externas Se presenta la precisión diagnóstica de detección general para radiólogos, la IA independiente y la referencia de decisión. Se proporcionan sensibilidad y especificidad para radiólogos (rojo), IA independiente (púrpura) y referencia de decisión (verde para la configuración ejemplar NT@97%+SN@98% y azul para configuraciones alternativas). Además, presentamos curvas ROC y AUROC para evaluar el rendimiento del sistema de IA en todo su rango operativo en el conjunto de prueba externo (n=82851; A) y en el subconjunto de datos para el que puede producir sus predicciones más confiables para la configuración ejemplar NT@97%+SN@98% (B). Las barras de error indican IC del 95 %. El enfoque de derivación de decisiones superó al radiólogo independiente en sensibilidad y especificidad, o en ambas, dependiendo de la configuración (A) al superar al radiólogo en todo el conjunto de predicciones confiables (B). Los valores de sensibilidad y especificidad resultantes para todos los estudios fueron similares o mayores que los del radiólogo solo, mientras que el 44,5–73,8 % de los estudios pudieron clasificarse de forma segura. IA = inteligencia artificial. AUC=área bajo la curva. AUROC=área bajo la característica operativa del receptor. NT = triaje normal. ROC = característica de funcionamiento del receptor. SN=red de seguridad.

mostrado. Los valores resultantes fueron similares o mayores que los obtenidos por el radiólogo sin ayuda, y entre el 42,1 y el 73,8 % de los estudios se pudieron clasificar de manera segura.

El rendimiento del sistema de IA en el subconjunto de datos para el que produjo sus predicciones más confiables es se muestra (figuras 3B, 4B). Con un AUROC de 0,982 (IC del 95 %: 0,977–0,986) en el conjunto de datos de la prueba interna y de 0,982 (0,978–0,985) en el conjunto de datos de la prueba externa, el rendimiento de la IA sistema superó el desempeño del radiólogo.

	Sensibilidad (95% IC)	Especificidad (95% CI)	Δ sensibilidad		Δ especificidad		Rendimiento de clasificación*
			Cambiar	valor p	Cambiar	valor p	
Datos de prueba interna							
Radiólogo	85·7% (83·6–87·9)	93·4% (93·1–93·7)	N / A	N / A	N / A	N / A	N / A
IA independiente	84·2% (82·4–85·8)	89·5% (89·0–89·9)	−1·5%	p=0·17	−3·9%	p<0·0001	89·5%
NT@0·95+SN@0·99	86·3% (84·1–88·0)	95·6% (95·3–95·9)	0·5%	p=0·43	2·2%	p<0·0001	71·1%
NT@0·95+SN@0·98	88·0% (86·1–89·8)	94·7% (94·4–95·0)	2·2%	p=0·0029	1·3%	p<0·0001	71·1%
NT@0·97+SN@0·99	88·0% (86·1–89·7)	94·8% (94·5–95·0)	2·2%	p=0·0001	1·4%	p<0·0001	60·7%
NT@0·98+SN@0·99	88·9% (87·1–90·7)	94·2% (93·9–94·5)	3·2%	p<0·0001	0·8%	p<0·0001	50·5%
NT@0·95+SN@0·97	88·9% (87·1–90·7)	93·8% (93·4–94·1)	3·2%	p<0·0001	0·4%	p=0·0097	71·1%
NT@0·99+SN@0·99	89·1% (87·3–90·9)	93·7% (93·4–94·0)	3·4%	p<0·0001	0·3%	p=0·0002	42·1%
NT@0·97+SN@0·98†	89·7% (87·9–91·3)	93·8% (93·6–94·1)	4·0%	p<0·0001	0·5%	p=0·0002	60·7%
NT@0·95+SN@0·95	90·0% (88·4–91·6)	91·7% (91·4–92·1)	4·3%	p<0·0001	−1·6%	p<0·0001	71·1%
NT@0·98+SN@0·98	90·6% (88·9–92·1)	93·3% (93·0–93·6)	4·9%	p<0·0001	−0·1%	p=0·33	50·5%
NT@0·97+SN@0·97	90·6% (88·8–92·1)	92·9% (92·6–93·2)	4·9%	p<0·0001	−0·5%	p=0·0006	60·7%
NT@0·99+SN@0·98	90·8% (89·1–92·4)	92·8% (92·5–93·1)	5·1%	p<0·0001	−0·6%	p<0·0001	42·1%
NT@0·98+SN@0·97	91·5% (89·9–93·1)	92·3% (92·0–92·7)	5·8%	p<0·0001	−1·1%	p<0·0001	50·5%
NT@0·97+SN@0·95	91·7% (90·2–93·2)	90·9% (90·5–91·3)	6·0%	p<0·0001	−2·5%	p<0·0001	60·7%
NT@0·99+SN@0·97	91·8% (90·2–93·3)	91·9% (91·5–92·2)	6·0%	p<0·0001	−1·5%	p<0·0001	42·1%
NT@0·98+SN@0·95	92·6% (91·2–94·1)	90·3% (89·9–90·7)	6·9%	p<0·0001	−3·1%	p<0·0001	50·5%
NT@0·99+SN@0·95	92·9% (91·3–94·3)	89·8% (89·4–90·2)	7·2%	p<0·0001	−3·5%	p<0·0001	42·1%
Datos de prueba externa							
Radiólogo	87·2% (85·6–88·7)	93·4% (93·2–93·6)	N / A	N / A	N / A	N / A	N / A
IA independiente	84·6% (83·3–85·9)	91·3% (91·1–91·5)	−2·6%	p=0·0019	−2·0%	p<0·0001	91·3%
NT@0·95+SN@0·99	86·8% (85·3–88·3)	95·9% (95·7–96·1)	−0·4%	p=0·45	2·5%	p<0·0001	73·8%
NT@0·95+SN@0·98	88·3% (86·9–89·7)	95·2% (95·0–95·4)	1·0%	p=0·06	1·8%	p<0·0001	73·8%
NT@0·97+SN@0·99	88·4% (87·0–89·7)	95·0% (94·9–95·2)	1·2%	p=0·0073	1·7%	p<0·0001	63·0%
NT@0·95+SN@0·97	89·0% (87·7–90·3)	94·5% (94·4–94·7)	1·8%	p=0·0011	1·2%	p<0·0001	73·8%
NT@0·98+SN@0·99	89·3% (87·8–90·6)	94·4% (94·2–94·6)	2·1%	p<0·0001	1·0%	p<0·0001	53·1%
NT@0·99+SN@0·99	89·6% (88·3–91·0)	94·0% (93·8–94·2)	2·4%	p<0·0001	0·6%	p<0·0001	44·5%
NT@0·97+SN@0·98†	89·8% (88·5–91·1)	94·3% (94·2–94·5)	2·6%	p<0·0001	1·0%	p<0·0001	63·0%
NT@0·95+SN@0·95	90·6% (89·3–91·7)	92·9% (92·7–93·1)	3·3%	p<0·0001	−0·4%	p<0·0001	73·8%
NT@0·97+SN@0·97	90·6% (89·4–91·9)	93·7% (93·5–93·9)	3·4%	p<0·0001	0·3%	p=0·0001	63·0%
NT@0·98+SN@0·98	90·7% (89·2–91·9)	93·7% (93·5–93·9)	3·5%	p<0·0001	0·3%	p<0·0001	53·1%
NT@0·99+SN@0·98	91·0% (89·7–92·2)	93·3% (93·1–93·5)	3·8%	p<0·0001	−0·1%	p=0·089	44·5%
NT@0·98+SN@0·97	91·5% (90·3–92·7)	93·0% (92·8–93·2)	4·2%	p<0·0001	−0·3%	p<0·0001	53·1%
NT@0·99+SN@0·97	91·8% (90·6–93·0)	92·6% (92·4–92·8)	4·6%	p<0·0001	−0·8%	p<0·0001	44·5%
NT@0·97+SN@0·95	92·1% (91·0–93·2)	92·1% (91·9–92·3)	4·9%	p<0·0001	−1·3%	p<0·0001	63·0%
NT@0·98+SN@0·95	93·0% (91·9–94·1)	91·4% (91·2–91·6)	5·8%	p<0·0001	−1·9%	p<0·0001	53·1%
NT@0·99+SN@0·95	93·3% (92·2–94·4)	91·0% (90·8–91·2)	6·1%	p<0·0001	−2·4%	p<0·0001	44·5%

Cada fila representa el punto de operación logrado en todos los estudios. Para la derivación de decisiones, cada fila se basa en dos umbrales que permitieron la categorización de los estudios que pasan por el proceso de derivación de decisiones en tres categorías, triaje normal, red de seguridad y derivación al radiólogo. La nomenclatura de configuración puede entenderse como NT@ que indica la sensibilidad del algoritmo en el conjunto de datos de validación para el punto operativo de triaje normal más SN@ que indica la especificidad del algoritmo en el conjunto de datos de validación para el punto operativo de la red de seguridad. La configuración del umbral y la selección de los puntos operativos en el conjunto de datos de validación se describen en el apéndice (pág. 3). Δ indica diferencia en sensibilidad y especificidad cuando se introduce AI. NT = triaje normal. SN=red de seguridad. *El rendimiento del triaje es la tasa de estudios etiquetados correctamente como normales (es decir, la fracción de estudios que podrían automatizarse). †Punto de funcionamiento ejemplar (NT@0·97+SN@0·98).

Tabla: Precisión diagnóstica y rendimiento de clasificación para radiólogos, IA independiente y referencia de decisiones en configuraciones seleccionadas para conjuntos de datos de pruebas internas y externas, donde cada fila representa un punto operativo resultante en todo el conjunto de datos

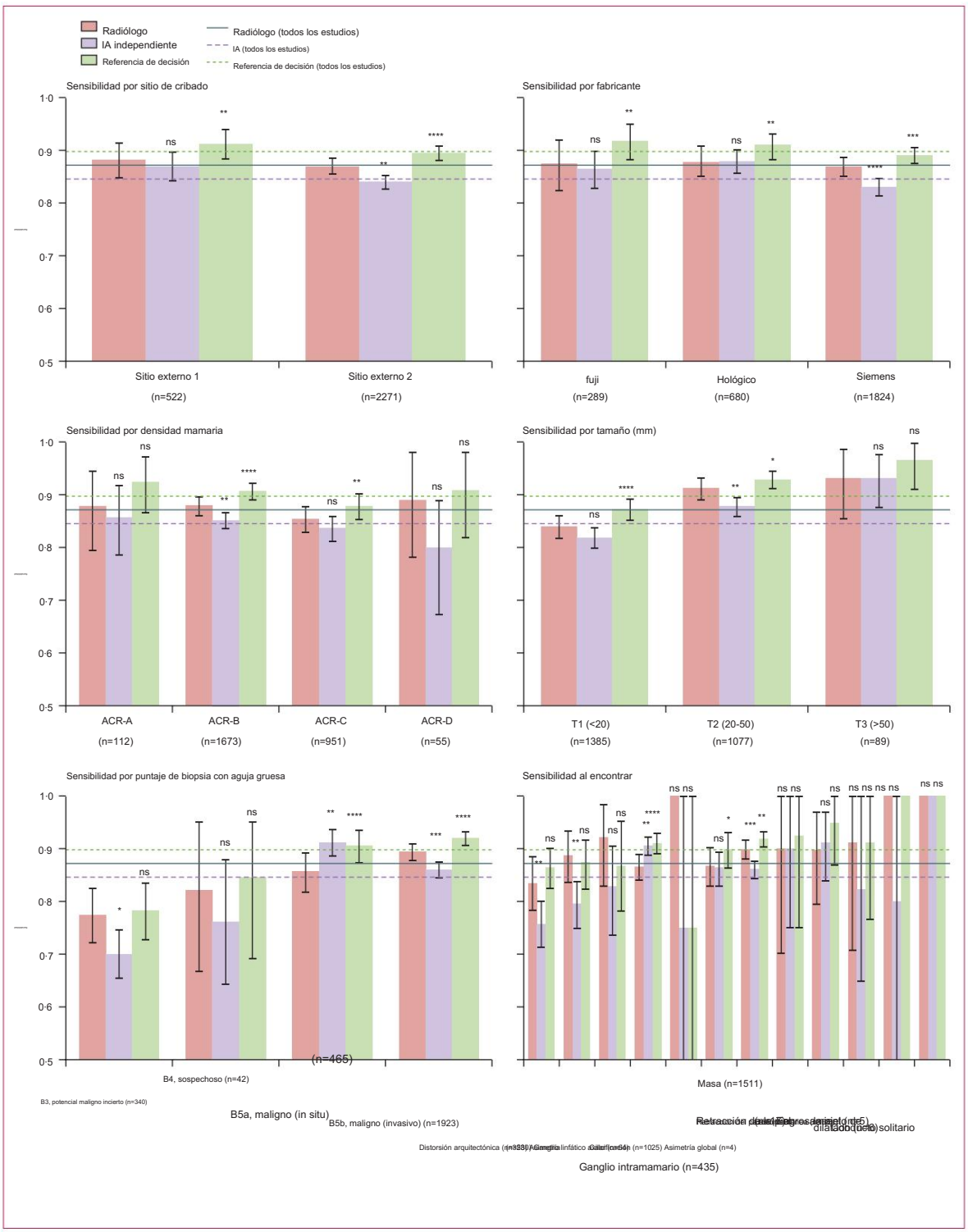

Figura 5: Desempeño de subgrupos en sensibilidad en configuración ejemplar en datos de prueba externa Las sensibilidades promedio para configuraciones ejemplares del enfoque de referencia de decisión (línea verde discontinua, NT@0·97+SN@0·98), son más altas que el promedio sensibilidad del radiólogo (línea roja continua) y sensibilidad promedio de IA independiente (línea púrpura discontinua, configuración como en la tabla). Los diagramas de barras muestran sensibilidades estratificadas en subgrupos relevantes. Los valores adjuntos están disponibles en el apéndice (pág. 9). IA = inteligencia artificial. ns=no significativo. NT = triaje normal. SN=red de seguridad. ****p≤0·001. ***p≤0·001. **p≤0·01. *p≤0·05.

Las sensibilidades promedio respectivas del sistema de IA independiente y del enfoque de referencia de decisiones en diferentes subgrupos se muestran mediante líneas horizontales punteadas y discontinuas (figura 5, para el conjunto de datos de prueba externa; apéndice p 12 para el conjunto de datos de prueba interna). El rendimiento difirió entre los diferentes subgrupos clínicos. Cuando la especificidad promedio se mantuvo constante, la sensibilidad promedio reducida para el IA independiente resultó en cambios negativos en la sensibilidad para varios subgrupos clínicamente relevantes. Por el contrario, la introducción de la derivación de decisiones resultó en cambios positivos significativos en la sensibilidad para varios subgrupos clínicamente relevantes. La introducción de la derivación de decisiones mejoró la capacidad del radiólogo para detectar lesiones malignas in situ e invasivas (conjunto de datos de prueba externa +4,9 % y +2,5 %, p≤0,0001 para ambos; conjunto de datos de prueba interna +3 ·8% y +4·1%, p=0·01 y p<0·0001). En los subgrupos estratificados por densidad mamaria, la referencia de decisión arrojó una sensibilidad significativamente mayor para los senos clasificados como ACR B (áreas dispersas de densidades fibroglandulares) y C (densos heterogéneos), que representan aproximadamente el 80 % de todas las mujeres examinadas.27 Los puntos porcentuales mejoraron de 1 · 8% a 4,5% en el conjunto de datos de prueba externa y 1,0% a 8,3% en el conjunto de datos de prueba interna. En subgrupos estratificados por hallazgos de imágenes, la derivación de decisiones mejoró la sensibilidad en varios subgrupos diferentes, incluso para masas y calcificaciones (conjunto de datos de prueba externa +1,9 % y +4,4 %, p=0,0013 y p<0,0001; conjunto de datos de prueba interna +3·7% y +5·1%, p<0·0001 para ambos). La derivación de decisiones mejoró la sensibilidad en todos los subgrupos estratificados por el tamaño de la lesión. A diferencia de la IA independiente, ningún subgrupo exhibió una disminución significativa en la sensibilidad cuando se utilizó el enfoque de referencia de decisiones. Los valores exactos se proporcionan en el apéndice (págs. 9, 12, 13).

En todas las configuraciones posibles, el algoritmo por sí solo generalizó desde un AUROC de 0,943 (95 % IC 0,937–0,949) en el conjunto de datos de validación (apéndice p 3) a un AUROC de 0,944 (0,939–0 ·950) en el conjunto de datos de prueba interna. Las configuraciones específicas diferían en términos de su generalizabilidad específica. El punto de operación de IA independiente se eligió para igualar la sensibilidad del radiólogo en el conjunto de datos de validación (apéndice p 3), a costa de una especificidad reducida en 4,5 puntos porcentuales (p<0,0001). En el conjunto de datos de la prueba interna, la sensibilidad y la especificidad de la compensación de la configuración independiente se desviaron hacia una reducción de la sensibilidad de 1,5 puntos porcentuales (p=0,17) y de la especificidad de 3,9 puntos porcentuales (p<0· 0001). El enfoque de referencia de decisión requiere que para una configuración elegida, las mejoras de sensibilidad y especificidad se mantengan en un conjunto de datos diferente y la evaluación algorítmica en el subconjunto confiable de estudios supere el desempeño del radiólogo en estos. En la configuración ejemplar, el enfoque de referencia de decisión mejoró la sensibilidad en 3,6 puntos porcentuales y la especificidad en 0,4 puntos porcentuales en el conjunto de datos de validación (apéndice p 3). Reutilizando la misma configuración, la sensibilidad mejoró en 4,0 puntos porcentuales y especificidad por 0·4 puntos porcentuales en el conjunto de datos de prueba interna. Para el subconjunto seguro de estudios, el algoritmo alcanzó un AUROC de 0,979 (IC del 95 %: 0,974–0,984) en el conjunto de datos de validación (apéndice p 5) y un AUROC de 0,982 (0,978–0 ·986) en el conjunto de datos de prueba interna (figura 3B), superando el desempeño del radiólogo sin ayuda en cada subconjunto seguro.

Para evaluar si el sistema de IA independiente y el enfoque de referencia de decisiones se pueden generalizar a nuevos sitios de detección que el algoritmo nunca antes había visto, el conjunto de datos externo se derivó de dos sitios de detección con diferentes radiólogos y mujeres. El sistema de IA independiente mantuvo un AUROC de 0·951 (0·947–0·955) en todas las configuraciones, y su punto de funcionamiento resultó en una reducción de 2,6 puntos porcentuales de sensibilidad y 2,0 puntos porcentuales de especificidad.

La decisión de derivación mantuvo cambios positivos al superar al radiólogo en los estudios de confianza (AUROC 0·982, 0·978–0·986; figura 4B), mejorando significativamente la sensibilidad en 2,6 puntos porcentuales y la especificidad en 1,0 puntos porcentuales (p <0·0001 para ambos).

Al estratificar por fabricante de dispositivo y sitio de detección, la IA independiente no pudo mantener la sensibilidad del radiólogo en todos los subgrupos, mientras que la referencia de decisión logró sensibilidades que no fueron significativamente más bajas o significativamente más altas que las del radiólogo sin ayuda (conjunto de datos de prueba externa, figura 5 y apéndice p 9; conjunto de datos de prueba interna, apéndice pp 12, 13). Además, la derivación de decisiones mantuvo o mejoró la especificidad del radiólogo en todos los fabricantes de dispositivos (apéndice, págs. 16, 17). En conjunto, el algoritmo y el enfoque de referencia de decisiones mostraron generalizabilidad en los ocho sitios de detección y dispositivos de mamografía diferentes de tres fabricantes diferentes.

Discusión

Nuestros resultados, basados en una evaluación de un sistema de IA que utiliza imágenes mamográficas recolectadas retrospectivamente de 4463 cánceres detectados por exámenes de detección y 100055 estudios normales de seguimiento comprobados, demuestran la aplicabilidad potencial de la IA a través de un enfoque de referencia de decisión, un triaje híbrido y cáncer. enfoque de detección. La simulación de este enfoque de derivación de decisiones mostró que la combinación de las fortalezas de los radiólogos y la IA podría generar mejoras notables en la sensibilidad y especificidad de los radiólogos individuales antes de la conferencia de consenso. Aunque el uso del sistema de IA en modo independiente en el conjunto de datos de pruebas externas mostró una reducción estadísticamente significativa de la sensibilidad del radiólogo en 2,6 puntos porcentuales y de la especificidad en 2,0 puntos porcentuales, los mismos modelos podrían usarse para colaborar con el radiólogo. en el modo de decisión-remisión. De hecho, la configuración ejemplar del sistema de IA dentro de un enfoque de decisión-referencia logró una mejora de la sensibilidad del radiólogo en 2,6 puntos porcentuales y la especificidad en 1,0 punto porcentual, al mismo tiempo que clasificaba automáticamente el 63,0 % de los estudios. Esto indica que el

La red de seguridad fue capaz de detectar cánceres que el primer lector pasó por alto y solo detectó el segundo lector.
La derivación de decisiones podría mejorar la sensibilidad y la especificidad generales, porque en el subconjunto de datos en los que el sistema de IA realizó predicciones, compuesto por cánceres detectados por detección y negativos comprobados durante el seguimiento, se obtuvo un AUROC de 0·982 que superó el rendimiento del radiólogo sin ayuda. logrado. Una serie de configuraciones alternativas del sistema de IA dentro de un enfoque de referencia de decisiones también logró un mejor rendimiento.

También confirmamos un rendimiento consistente y mejorado del enfoque de derivación de decisiones en subgrupos clínicamente relevantes, incluidos aquellos que se presentan como casos desafiantes para los radiólogos. La sensibilidad también fue consistente en tres fabricantes de dispositivos diferentes y ocho sitios de detección diferentes. Es de destacar que un modelo de IA, si se implementa en la práctica clínica, también tiene el potencial de mejorarse aún más al recibir capacitación sobre los nuevos datos entrantes, lo que garantiza que el rendimiento en todos los subgrupos no se degrade.

El enfoque de referencia de decisiones permitiría que los programas de detección trabajen iterativamente para automatizar más decisiones de detección dentro de un marco seguro, en lugar de convertirse en un sistema de IA totalmente automatizado sin supervisión humana. La literatura existente sobre la precisión de los sistemas de IA no respalda la implementación de aplicaciones independientes en la práctica clínica.9,28 Una revisión sistemática de la literatura publicada encontró que 34 (94 %) de 35 estudios de sistemas de IA fueron menos precisos que un solo radiólogo , mientras que los pocos estudios pequeños que mostraron una mayor precisión de un sistema independiente tenían un alto riesgo de sesgo y tenían poca generalización al contexto clínico.9 En modo independiente, nuestra IA logró una sensibilidad del 84,6 % y una especificidad del 91· 3% en datos externos, también con un desempeño menos preciso que el radiólogo único promedio. Existen claras advertencias que dificultan la adopción de un sistema independiente. En entornos de baja prevalencia de cáncer (es decir, detección), la variabilidad de los valores predictivos positivos entre los radiólogos da como resultado falsos positivos, lo que requiere recursos adicionales para la revisión por consenso y las pruebas de diagnóstico.29,30 La IA completamente automatizada no mejora este desafío; Las predicciones ambiguas de IA aún darían como resultado una gran cantidad de falsos positivos y una mayor carga de trabajo. A diferencia de los enfoques de IA independientes, el enfoque de referencia de decisiones solo toma decisiones sobre un subconjunto de exámenes con un alto grado de precisión. Con una mayor mejora del modelo, se espera que aumente esta fracción de decisiones precisas.

El enfoque de derivación de decisiones se diferencia aún más del modelo de conjunto y los enfoques de clasificación independientes porque combina la clasificación automatizada de casos normales y la derivación de decisiones que integra una red de seguridad para la predicción positiva de casos; pero con respecto a la red de seguridad, el modelo intencionalmente no proporciona acceso directo a las predicciones de los exámenes referidos al usuario para evitar posibles sesgos engañosos. En la práctica, modelo de trabajo superior

Las predicciones se presentarían como informes normales precargados y las predicciones positivas del modelo como advertencias de la red de seguridad, y solo se mostrarían si un radiólogo asignó una puntuación BI-RADS inferior a 3. Por lo tanto, una evaluación definitiva del rendimiento general requiere decisiones finales del radiólogo después de sugerencias algorítmicas, manteniendo la supervisión humana final.

Reconocemos las limitaciones inherentes a la evaluación del enfoque de decisión-remisión en un entorno retrospectivo. El conjunto de datos retrospectivos excluyó los casos que no tuvieron un seguimiento normal dentro de los 4·5 años posteriores a la selección. Creemos que este es un período generoso para capturar una cohorte diversa de mujeres con diferentes prácticas de detección, por ejemplo, incluidas aquellas que podrían no cumplir con las pautas de detección bienales. Sin embargo, este enfoque podría resultar en la exclusión de las mujeres que asistían a su última cita de detección a los 69 años, o que abandonaron la detección por completo.

Nuestro análisis requería la suposición de que las predicciones confiables se realizan automáticamente. Los sistemas de IA para el rendimiento y las tareas críticas para la seguridad deben probarse exhaustivamente antes de tomar decisiones automatizadas. El papel del radiólogo sigue siendo fundamental para el enfoque de derivación de decisiones que proponemos. Sin embargo, esta fue una simulación que no tuvo en cuenta la interacción humano-IA, lo que impidió una evaluación directa de cómo las recomendaciones generadas por IA influyen en la toma de decisiones de los radiólogos. Concretamente, la simulación asumió que el radiólogo no corregiría ninguna de las sugerencias algorítmicas, de modo que se asumió que los informes normales precargados se aceptaban incluso si esto conducía a un cáncer pasado por alto, y se aceptaban las advertencias de la red de seguridad incluso si eran falsas. positivos Con la supervisión humana, las predicciones de IA erróneas pero corregidas solo pueden conducir a una mejora adicional de las métricas de detección. Para predicciones correctas y aceptadas, nuestros hallazgos reflejan los mejores resultados posibles. Para las predicciones de IA correctas pero no aceptadas, el algoritmo no puede ser directamente responsable, pero la educación cuidadosa y el monitoreo de las predicciones del triaje normal, la red de seguridad y los radiólogos deberían ser obligatorios para que los proveedores de IA no repitan las trampas de los sistemas de detección asistidos por computadora. 31 Con predicciones de IA más precisas y confiables y estudios referidos (con una prevalencia de cáncer similar a la de la población general), el enfoque de decisión de referencia es prometedor. En última instancia, solo las evaluaciones prospectivas de la interacción humana y de IA en una cohorte completamente representativa de mujeres que asisten a la selección podrían proporcionar información directa sobre la influencia del enfoque de derivación de decisiones en la toma de decisiones del radiólogo.

Otra limitación de este estudio es que evaluó el desempeño de un solo lector antes de la conferencia de consenso utilizando el enfoque de decisión-remisión. Un enfoque para reducir aún más la carga de trabajo es hacer que ambos lectores en el entorno de doble lector utilicen el enfoque de referencia de decisión. Con un

del 50 % logrado para cada lector, este enfoque daría como resultado una reducción total de la carga de trabajo de más del 100 % (de un 200 %), superando lo que podría lograr una solución de IA independiente al reemplazar un lector (100 %). Sin embargo, comprender los efectos más amplios de aplicar el mismo enfoque a dos lectores es una importante tarea de investigación futura que también debe incluir información sobre los cánceres de intervalo.

Este estudio ha proporcionado evidencia para continuar en el camino hacia la adopción clínica generalizada y segura de sistemas basados en IA para el cribado de mamas. El enfoque de derivación de decisiones aprovecha las fortalezas tanto del radiólogo como del algoritmo de IA, lo que demuestra que se pueden realizar mejoras en la sensibilidad y la especificidad que superan las del radiólogo individual y el sistema de IA independiente, incluso si se utiliza el mismo algoritmo subyacente. Este enfoque tiene el potencial de mejorar la precisión de la detección de los radiólogos, se adapta a los requisitos (heterogéneos) de la detección y podría permitir la reducción de la carga de trabajo a través de estudios normales de clasificación, sin descartar la supervisión final de los radiólogos.

Colaboradores

Todos los autores contribuyeron a la concepción, el diseño o ambos elementos del estudio y tuvieron acceso a todos los conjuntos de datos sin procesar en todo momento. CL, MB y SB participaron en la adquisición y conservación de los datos y han verificado los datos subyacentes. Verificar la consistencia de los datos requeridos al relacionar los datos sin procesar con los datos preprocesados; los autores supervisores y externos (KP y LU) tenían la capacidad de solicitar verificación en cualquier momento. CL y SB desarrollaron la red neuronal artificial e hicieron los análisis. MB supervisó la anotación de los estudios utilizados para el desarrollo de IA. KP y LU proporcionaron orientación conceptual. CL hizo el análisis estadístico, que fue revisado por SB.

El manuscrito fue escrito por CL y DB, y los demás autores brindaron apoyo editorial. Todos los autores leyeron y aprobaron el manuscrito final.

Declaración de intereses CL,

MB, SB y DB son empleados de Vara, el financiador del estudio. LU es asesor médico de Vara (MX Healthcare), orador y miembro del consejo asesor de Bayer Healthcare, y recibió una subvención de investigación de Siemens Healthcare fuera del trabajo presentado. KP es el principal asesor médico de Vara (MX Healthcare) y recibió pagos por actividades no relacionadas con el presente artículo, incluidas conferencias y servicios en las oficinas de oradores y por gastos de viaje, alojamiento y reuniones no relacionados con las actividades enumeradas por la Sociedad Europea de Mama. Imaging (curso educativo de resonancia magnética y reunión científica anual), el IDKD 2019 (curso educativo) y Siemens Healthineers.

Intercambio de

datos La información adicional relacionada con este estudio está disponible previa solicitud al autor correspondiente. El código utilizado para procesar los datos sin procesar y desarrollar el modelo está estrechamente integrado con un sistema de producción comercial y, por lo tanto, no puede publicarse. Sin embargo, la procedencia del modelo ejemplar utilizado para este trabajo se describe en el apéndice, que se puede utilizar junto con marcos de aprendizaje profundo de código abierto como TensorFlow o PyTorch. La principal contribución de esta publicación es cómo utilizar y evaluar cualquier modelo suficientemente preciso para la clasificación del cáncer de mama en mamografías para allanar el camino hacia la aplicabilidad clínica. Por lo tanto, todos los detalles de la evaluación están cuidadosamente documentados y la parte de evaluación del código junto con los datos para reproducir figuras y tablas están disponibles en https://github. com/vara-ai/decision-referral.

Expresiones de gratitud

Reconocemos y agradecemos a las ocho unidades alemanas de detección de mama participantes por brindar acceso a los estudios de mamografía. y datos de pacientes anónimos, y a los radiólogos involucrados en anotar las imágenes de mamografía. Nuestro agradecimiento también a Dipti Ganeriwala por diseñar la figura 1 y la figura 2. Jack Dunger, Zacharias V Fisches, Thijs Kooi, Dominik Schüler, Benjamin Strauch y Vilim Štih (orden alfabético) por discusiones fructíferas, ideas, análisis de apoyo e ingeniería que directa o indirectamente hizo posible este trabajo. Los autores agradecen a Joanne Chin por su ayuda en la edición de este manuscrito. El trabajo inicial en el desarrollo del algoritmo fue patrocinado por una subvención pública Pro FIT de Investitionsbank Berlin (número de subvención 10166923) y una subvención del programa Eurostars Horizon 2020 de la Comisión Europea (01QE2002).

Referencias

1 Litjens G, Kooi T, Bejnordi BE, et al. Una encuesta sobre el aprendizaje profundo en el análisis de imágenes médicas. Análisis de Imágenes Médicas 2017; 42: 60–88.

2 Kim HE, Kim HH, Han BK, et al. Cambios en la detección de cáncer y recuperación de falsos positivos en mamografías usando inteligencia artificial: un estudio retrospectivo de múltiples lectores. Lancet Digit Salud 2020; 2: e138–48.

3 McKinney SM, Sieniek M, Godbole V, et al. Evaluación internacional de un sistema de IA para el cribado del cáncer de mama. Naturaleza 2020; 577: 89–94.

4 Ribli D, Horváth A, Unger Z, Pollner P, Csabai I. Detección y clasificación de lesiones en mamografías con aprendizaje profundo. representante científico 2018; 8: 4165.

5 Rodríguez-Ruiz A, Krupinski E, Mordang JJ, et al. Detección de cáncer de mama con mamografía: efecto de un sistema de apoyo de inteligencia artificial. Radiología 2019; 290: 305–14.

6 Salim M, Wåhlin E, Dembrow K, et al. Evaluación externa de 3 algoritmos comerciales de inteligencia artificial para la evaluación independiente de mamografías de detección. JAMA Oncol 2020; 6: 1581–88.

7 Schaffter T, Buist DSM, Lee CI, et al. Evaluación de la inteligencia artificial combinada y la evaluación del radiólogo para interpretar las mamografías de detección. Red JAMA Abierta 2020; 3: e200265.

8 Wu N, Phang J, Park J, et al. Las redes neuronales profundas mejoran el desempeño de los radiólogos en la detección del cáncer de mama. IEEE Transact Med Imag 2019; 39: 1184–94.

9 Freeman K, Geppert J, Stinton C, et al. Uso de inteligencia artificial para el análisis de imágenes en programas de detección de cáncer de mama: revisión sistemática de la precisión de la prueba. BMJ2021 ; 374: n1872.

10 Dembrower K, Wåhlin E, Liu Y, et al. Efecto del triaje basado en inteligencia artificial de las mamografías de detección del cáncer de mama en la detección del cáncer y la carga de trabajo del radiólogo: un estudio de simulación retrospectivo. Lancet Digit Salud 2020; 2: e468–74.

11 Kyono T, Gilbert FJ, van der Schaar M. Mejora de la eficiencia del flujo de trabajo para mamografía mediante el aprendizaje automático. J Am College Radiol 2020; 17: 56–63.

12 Raya-Povedano JL, Romero-Martín S, Elías-Cabot E, Gubern-Mérida A, Rodríguez-Ruiz A, Álvarez-Benito M. Estrategias basadas en IA para reducir la carga de trabajo en el tamizaje de cáncer de mama con mamografía y tomosíntesis: una evaluación retrospectiva. Radiología 2021; 1: 203555.

13 Yala A, Schuster T, Miles R, Barzilay R, Lehman C. Un modelo de aprendizaje profundo para clasificar las mamografías de detección: un estudio de simulación. Radiología 2019; 293: 38–46.

14 Balta C, Rodríguez-Ruiz A, Mieskes C, Karssemeijer N, Heywang-Köbrunner S. Pasar de la lectura doble a la simple para los exámenes de detección etiquetados como probablemente normales por la IA: ¿cuál es el impacto?: SPIE 11513 15.º Taller internacional sobre imágenes mamarias (IWBI2020); 22 de mayo de 2020 (115130D).

15 Lång K, Dustler M, Dahlblom V, Åkesson A, Andersson I, Zackrisson S. Identificación de mamografías normales en una gran población de detección mediante inteligencia artificial. Europ Radiol 2021; 31: 1687–92.

16 Leibig C, Allken V, Ayhan MS, Berens P, Wahl S. Aprovechamiento información de incertidumbre de redes neuronales profundas para la detección de enfermedades. representante científico 2017; 7: 17816.

17 Lakshminarayanan B, Pritzel A, Blundell C. Estimación de incertidumbre predictiva simple y escalable usando conjuntos profundos. Sistema de proceso de información neuronal avanzado 2017; 30: 1.

18 James G, Witten D, Hastie T, Tibshirani R. Una introducción al aprendizaje estadístico. Nueva York: Springer, 2013.

19 Geirhos R, Jacobsen JH, Michaelis C, et al. Aprendizaje de atajos en redes neuronales profundas. NAT Machine Intellig 2020; 2: 665–73.

20 Pinsky PF, Gallas B. Diseños enriquecidos para evaluar el desempeño discriminatorio: análisis de sesgo y varianza. Stat Med 2012; 31: 501–15.

21 Mansurnia MA, Altman DG. Ponderación de probabilidad inversa. BMJ 2016; 352: i189.

22 Amendoeira I, Perry N, Broeders M, et al. Directrices europeas para la garantía de calidad en el cribado y diagnóstico del cáncer de mama. Bruselas: Comisión Europea, 2013.

23 Forester ND, Lowes S, Mitchell E, Twiddy M. Lesiones mamarias de alto riesgo (B3): ¿cuál es la incidencia de malignidad para los subtipos de lesiones individuales? Una revisión sistemática y metanálisis. Eur J Surg Oncol 2019; 45: 519–27.

24 Lee A, Anderson N, Carder P, et al. Directrices para procedimientos de diagnóstico no quirúrgicos y notificación en la detección del cáncer de mama. Londres: The Royal College of Pathologists, 2016.

25 Efron B, Tibshirani RJ. Una introducción al bootstrap. Boca Ratón: CRC Press, 1994.

26 Sickles EA, D'Orsi CJ, Bassett LW, et al. Mamografía ACR BI RADS®. En: ACR BI-RADS® Atlas, Sistema de datos e informes de imágenes mamarias. Reston, VA: Colegio Americano de Radiología, 2013.

27 Winkler NS, Raza S, Mackesy M, Birdwell RL. Densidad mamaria: Implicaciones clínicas y métodos de evaluación. Radiografías 2015; 35: 316–24.

28 Aggarwal R, Sounderajah V, Martin G, et al. Precisión diagnóstica del aprendizaje profundo en imágenes médicas: una revisión sistemática y un metanálisis. medicina digital NPJ 2021; 4: 65.

29 Hofvind S, Ponti A, Patnick J, et al. Resultados falsos positivos en Cribado mamográfico para el cáncer de mama en Europa: una revisión de la literatura y un estudio de los programas de cribado de servicios. J Med Cribado 2012; 19 (suplemento 1): 57–66.

30 Lazarus E, Mainiero MB, Schepps B, Koelliker SL, Livingston LS. Léxico BI-RADS para ecografía y mamografía: variabilidad interobservador y valor predictivo positivo. Radiología 2006; 239: 385–91.

31 Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL. Precisión diagnóstica de la mamografía de detección digital con y sin detección asistida por computadora. JAMA Intern Med 2015; 175: 1828-1837.