Las redes neuronales profundas mejoran la de los radiólogos Desempeño en la detección del cáncer de mama

nan wu, Jason Phang, Parque Jungkyu, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisł aw Jastrze ̨bski, Thibault Févry, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikh, Sushma Gaddam, Leng Leng Young Lin, Kara Ho, Joshua D. Weinstein, Beatriu Reig, Yiming Gao, Hildegard Toth, Kristine Pysarenko, Alana Lewin, Jiyon Lee, Krystal Airola, Eralda Mema, Stephanie Chung, Esther Hwang, Naziya Samreen, S. Gene Kim, Laura Heacock, Linda Moy, Kyunghyun Cho, y Krzysztof J. Geras

Abstracto —Presentamos una red neuronal convolucional profunda para la clasificación de exámenes de detección de cáncer de mama, entrenada y evaluada en más de 200 000 exámenes (más de 1 000 000 de imágenes). Nuestra red logra un AUC de 0,895 en la predicción de la presencia de cáncer de mama, cuando se prueba en la población de cribado. Atribuimos la alta precisión a algunos avances técnicos. 1) Nuestra red novedosa arquitectura de dos etapas y procedimiento de entrenamiento, que nos permite usar una red de nivel de parche de alta capacidad para aprender de etiquetas de nivel de píxel junto con una red que aprende de etiquetas macroscópicas de nivel de seno. 2) Una red personalizada basada en ResNet utilizada como elemento básico de nuestro modelo, cuyo equilibrio de profundidad y ancho está optimizado para imágenes médicas de alta resolución. 3) Capacitación previa de la red en la detección de la clasificación BI-RADS, una tarea relacionada con etiquetas más ruidosas. 4) Combinar múltiples vistas de entrada de manera óptima entre un número de opciones posibles. Para validar nuestro modelo, llevamos a cabo un estudio de lectores con 14 lectores, cada uno leyendo 720 exámenes de mamografía de detección, y demostramos que nuestro modelo es tan preciso como radiólogos experimentados cuando se les presentan los mismos datos. También mostramos que un modelo híbrido, que promedia la probabilidad de malignidad predicha por un radiólogo con una predicción de nuestra red neuronal, es más preciso que cualquiera de los dos por separado. Para comprender mejor nuestros resultados, llevamos a cabo un análisis exhaustivo del rendimiento de nuestra red en diferentes subpoblaciones de la población de detección, el diseño del modelo, el procedimiento de entrenamiento, los errores y las propiedades de sus representaciones internas. Nuestros mejores modelos están disponibles públicamente en https://github.com/nyukat/breast_cancer_ classifier.

Índice Términos—Aprendizaje profundo, redes neuronales convolucionales profundas, detección de cáncer de mama, mamografía.

Manuscrito recibido el 19 de junio de 2019; revisado el 16 de septiembre de 2019; aceptado el 28 de septiembre de 2019. Fecha de publicación 7 de octubre de 2019; fecha de la versión actual 1 de abril de 2020. Este trabajo fue apoyado en parte por los Institutos Nacionales de Salud bajo la subvención R21CA225175 y la subvención P41EB017183. (Autor para correspondencia: Krzysztof J. Gerás.)

N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang y T. Févry trabajan en el Center for Data Science, New York University, New York, NY 10011 EE. UU.

M. Zorin trabajó en el Instituto Courant de Ciencias Matemáticas de la NYU, Universidad de Nueva York, Nueva York, NY 10011, EE. UU. Ahora trabaja en el Departamento de Ciencias de la Computación y Tecnología de la Universidad de Cambridge, Cambridge CB3 0FD, Reino Unido.

S. Jastrze ̨bski está en la Facultad de Matemáticas e Información Technologies, Universidad Jagellónica, 30-348 Cracovia, Polonia.

J. Katsnelson, E. Kim, S. Wolfson, U. Parikh, S. Gaddam, LLY Lin, JD Weinstein, K. Airola, E. Mema, S. Chung, E. Hwang y N. Samreen están en el Departamento de Radiología, Facultad de Medicina, Universidad de Nueva York, Nueva York, NY 10016 EE. UU.

K. Ho está en el SUNY Downstate College of Medicine, Nueva York, Nueva York 11203 EE. UU.

B. Reig, Y. Gao, H. Toth, K. Pysarenko, A. Lewin, J. Lee y L. Heacock trabajan en el Departamento de Radiología, Escuela de Medicina, Universidad de Nueva York, Nueva York, NY 10016 EE. UU., y también con el Perlmutter Cancer Center, NYU Langone Health, New York, NY 10016 USA.

SG Kim y L. Moy están en el Departamento de Radiología, Escuela de Medicina, Universidad de Nueva York, Nueva York, NY 10016 EE. UU., en el Centro de Cáncer Perlmutter, NYU Langone Health, Nueva York, NY 10016 EE. UU., y también en el Centro para la Innovación e Investigación en Imágenes Avanzadas, NYU Langone Health, Nueva York, NY 10016 EE. UU.

K. Cho trabaja en el Center for Data Science, New York University, New York, NY 10011 USA, y también en el Courant Institute of Mathematical Sciences, New York University, New York, NY 10012 USA.

KJ Geras está en el Departamento de Radiología, Escuela de Medicina, Universidad de Nueva York, Nueva York, NY 10016 EE. UU., en el Centro de Ciencia de Datos, Universidad de Nueva York, Nueva York, NY 10011 EE. UU., y también en el Centro de Innovación en Imágenes Avanzadas and Research, NYU Langone Health, Nueva York, NY 10016 EE. UU. (correo electrónico: kjgeras@nyu.edu).

Este artículo tiene material descargable complementario disponible en http://ieeexplore.ieee.org, proporcionado por los autores.

Las versiones en color de una o más de las figuras de este artículo están disponibles. en línea en http://ieeexplore.ieee.org.

Identificador de objeto digital 10.1109/TMI.2019.2945514

I. INTRODUCCIÓN

El cáncer de MAMA es el segundo cáncer relacionado causa de muerte entre las mujeres en los EE. UU. En 2014, se realizaron más de 39 millones de mamografías de detección y diagnóstico en los EE. UU. Se estima que en 2015 232.000 mujeres fueron diagnosticadas con cáncer de mama y aproximadamente 40.000 fallecieron a causa de este [1]. Aunque la mamografía es la única prueba de diagnóstico por la imagen que ha reducido la mortalidad por cáncer de mama [2]–[4], se ha debatido sobre los daños potenciales de las pruebas de detección, incluidos los retiros de falsos positivos y las biopsias falsas positivas asociadas. La gran mayoría del 10% al 15% de las mujeres a las que se les pide regresar después de una mamografía de detección no concluyente se someten a otra mamografía y/o ecografía para aclaraciones. Después de los exámenes de imágenes adicionales, muchos de estos hallazgos se determinan como benignos y solo se recomienda que entre el 10 y el 20 % se somet

trabajo adicional. Entre estos, solo el 20-40% arrojan un diagnóstico de cáncer [5]. Evidentemente, existe una necesidad insatisfecha de cambiar el equilibrio de la detección sistemática del cáncer de mama hacia más beneficios y menos daños.

Los radiólogos utilizan habitualmente la detección asistida por computadora (CAD) tradicional en mamografía para ayudar con la interpretación de imágenes, a pesar de que los estudios multicéntricos muestran que estos programas CAD no mejoran su rendimiento diagnóstico [6].
Estos programas suelen utilizar funciones artesanales para marcar sitios en una mamografía que se diferencian del tejido normal. El radiólogo decide si recuerda estos hallazgos, determinando la importancia clínica y la capacidad de acción. Los desarrollos recientes en el aprendizaje profundo [7], en particular, las redes neuronales convo lucionales profundas (CNN) [8]–[12], abren posibilidades para crear una nueva generación de herramientas similares a CAD.

Este documento realiza varias contribuciones técnicas hacia el objetivo de desarrollar redes neuronales para ayudar a los radiólogos a interpretar los exámenes de detección del cáncer de mama. (i) Presentamos una nueva red neuronal de dos etapas para incorporar información global y local con un procedimiento de entrenamiento adecuado. Esto nos permitió utilizar una red de nivel de parche de muy alta capacidad para aprender de etiquetas de nivel de píxel junto con una red que aprende de etiquetas macroscópicas de nivel de pecho. Con esta estrategia, nuestro modelo no solo logra un rendimiento humano-competitivo, sino que también produce mapas de calor interpretables que indican ubicaciones de hallazgos sospechosos. Además, mostramos la utilidad de las etiquetas a nivel de píxel incluso en un régimen en el que tenemos muchas etiquetas a nivel de imagen. (ii) Demostramos la viabilidad de entrenar y evaluar la red con más de 1 000 000 de imágenes mamográficas de alta resolución, un conjunto de datos extremadamente grande en imágenes médicas, no solo para la detección del cáncer de mama. Esto tiene un valor significativo tanto para informar futuras prioridades de diseño de investigación como para mostrar una prueba de concepto y el valor de prueba de este enfoque. Además, realizamos un cuidadoso análisis de errores de nuestras predicciones e identificamos patrones que nuestra red no pudo capturar, lo que informará los futuros diseños de arquitectura. (iii) Para usar como componente básico de nuestra red, proponemos una variante novedosa de ResNet diseñada específicamente para imágenes médicas, que tiene un equilibrio de profundidad y ancho que permite que el modelo procese una imagen muy grande mientras mantiene un consumo de memoria razonable. . (iv) Evaluamos la utilidad de preentrenar la red usando una tarea relacionada con un resultado más ruidoso (detección de la clasificación BI RADS) y encontramos que es una parte muy importante de la canalización que mejora notablemente el rendimiento de nuestros modelos. Esto es de particular importancia en el campo de las imágenes médicas, donde la mayoría de los conjuntos de datos son pequeños. (v) Evaluamos varias formas de combinar información de diferentes vistas mamográficas dentro de una sola red neuronal. Los resultados de este análisis también son valiosos para una audiencia más amplia, incluidos los radiólogos, particularmente en relación con el margen de rendimiento entre modelos entrenados en un subconjunto de las vistas.

No conocemos ningún análisis previo como este, aunque es común que las tareas de imágenes médicas tengan múltiples entradas. (vi) Hemos puesto a disposición el código y los pesos de nuestros mejores modelos en https://github.com/nyukat/breast_cancer_classifier.
Con esta aportación, grupos de investigación que están trabajando en la mejora de la mamografía de cribado, que pueden no tener acceso a un gran conjunto de datos de entrenamiento como el nuestro, podrán usar directamente nuestro modelo en su investigación o usar nuestros pesos preentrenados como una inicialización para entrenar modelos con menos datos. Al hacer públicos nuestros modelos, invitamos a otros grupos a validar nuestros resultados y probar su robustez ante cambios en la distribución de datos.

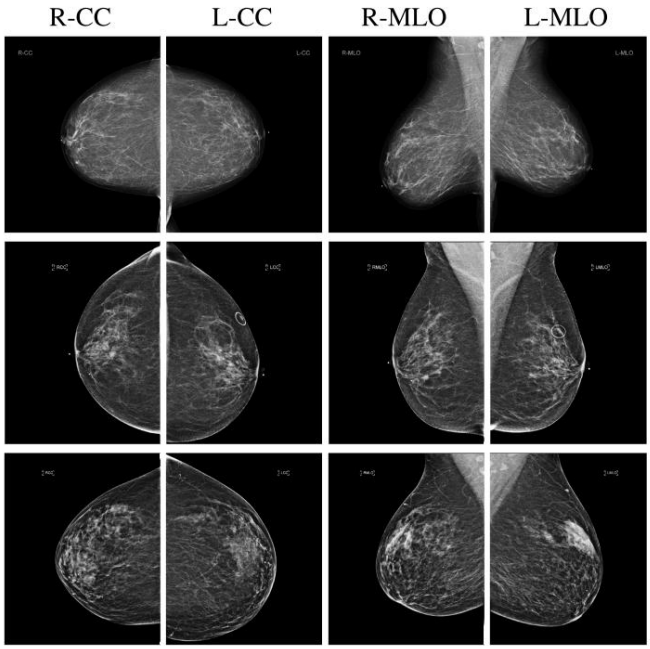

Fig. 1. Ejemplos de exámenes de detección de cáncer de mama. Primera fila: ambas mamas sin hallazgos; segunda fila: mama izquierda sin hallazgos y mama derecha con hallazgo maligno; tercera fila: mama izquierda con hallazgo benigno y mama derecha sin hallazgos.

II. DATOS

Nuestro estudio retrospectivo fue aprobado por nuestra junta de revisión institucional y cumplió con la Ley de Portabilidad y Responsabilidad del Seguro Médico. Se renunció al consentimiento informado. Este conjunto de datos1 es una versión más grande y cuidadosamente seleccionada de un conjunto de datos utilizado en nuestro trabajo anterior [14], [15].
El conjunto de datos incluye 229 426 exámenes de mamografía de detección digital (1 001 093 imágenes) de 141 473 pacientes. Cada examen contiene al menos cuatro imágenes,2 correspondientes a las cuatro vistas estándar utilizadas en la mamografía de detección: R-CC (craneocaudal derecha), L-CC (craneocaudal izquierda), R-MLO (oblicua media lateral derecha) y L-MLO (oblicua mediolateral izquierda). Las imágenes del conjunto de datos provienen de cuatro tipos de escáneres: Mammomat Inspiration (22,81 %), Mammomat Novation DR (12,65 %), Lorad Selenia (40,92 %) y Selenia Dimensions (23,62 %). En la figura 1 se muestran algunos ejemplos de exámenes .

Para extraer etiquetas que indiquen si se encontró que cada seno de la paciente tenía hallazgos malignos o benignos al final del proceso de diagnóstico, nos basamos en los informes patológicos de las biopsias. Tenemos 5,832 exámenes con al menos una biopsia realizada dentro de los 120 días posteriores a la mamografía de detección. Entre estos, las biopsias confirmaron hallazgos malignos para

1Los detalles de sus estadísticas y cómo se extrajeron se pueden encontrar en un informe técnico separado [13].

2Algunos exámenes contienen más de una imagen por vista, ya que es posible que los tecnólogos necesiten repetir una imagen o proporcionar una vista complementaria para obtener una imagen completa del seno en un examen de detección.

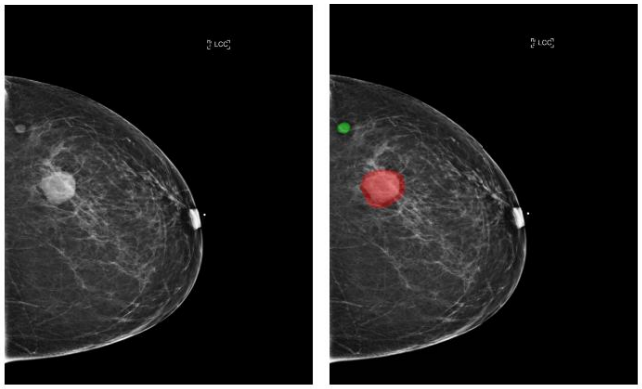

Figura 2. Ejemplo de segmentación realizada por un radiólogo.
Izquierda: la imagen original. Derecha: la imagen con lesiones que requieren una biopsia resaltada. El hallazgo maligno se resalta con rojo y el hallazgo benigno con verde.

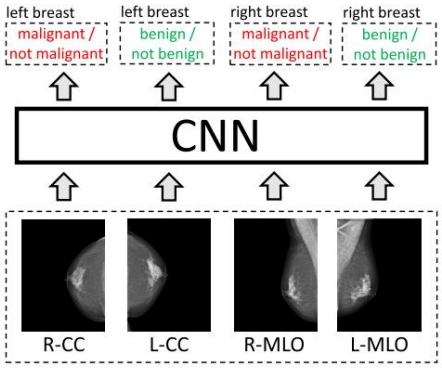

Fig. 3. Una representación esquemática de cómo formulamos la clasificación del examen de cáncer de mama como una tarea de aprendizaje. La principal tarea que pretendemos que aprenda el modelo es la clasificación maligno/no maligno. La tarea de clasificación benigna/no benigna se utiliza como tarea auxiliar de regularización de la red.

TABLA I
NÚMERO DE MAMA CON HALLAZGOS MALIGNOS Y BENIGNOS EN BASE A LAS ETIQUETAS EXTRAÍDAS DE LA PATOLOGÍA INFORMES , DESGLOSADOS SEGÚN SI LOS HALLAZGOS FUERON VISIBLES U OCULTOS

	malignant		benign	
	visible	occult	visible	occult
training	750	107	2,586	2,004
validation	51	15	357	253
test	54	8	215	141
overall	855 (86.8%)	130 (13.2%)	3,158 (56.84%)	2,398 (43.16%)

985 (8,4%) mamas y hallazgos benignos para 5.556 (47,6%) mamas. 234 (2,0%) mamas tenían hallazgos tanto malignos como benignos. Para los exámenes de detección restantes que no coincidieron con una biopsia, asignamos etiquetas correspondientes a la ausencia de hallazgos malignos y benignos en ambos senos.

Para todos los exámenes emparejados con biopsias, solicitamos a un grupo de radiólogos (provistos de los informes de patología correspondientes) que indicaran retrospectivamente la ubicación de las lesiones biopsiadas a nivel de píxel. En la figura 2 se muestra un ejemplo de dicha segmentación. Encontramos que aproximadamente el 32,8 % de los exámenes estaban ocultos en la mamografía, es decir, las lesiones que se biopsiaron no eran visibles en la mamografía, incluso retrospectivamente, y se identificaron mediante otras modalidades de imagen: ecografía o resonancia magnética. Ver Tabla I para más detalles.

tercero CNNS PROFUNDAS PARA LA CLASIFICACIÓN DEL CÁNCER

Dado que algunas mamas contienen hallazgos tanto malignos como benignos, formulamos la clasificación del cribado del cáncer de mama como una tarea de aprendizaje utilizando el marco de aprendizaje multitarea [16]. Es decir, para cada seno, asignamos dos etiquetas binarias: la ausencia/presencia de hallazgos malignos en un seno (indicado por yR,m y yL,m) y la ausencia/presencia de hallazgos benignos en un seno (indicado por yR ,b y yL,b). Con los senos izquierdo y derecho, cada examen tiene un total de cuatro etiquetas binarias. Nuestro objetivo es producir cuatro predicciones correspondientes a las cuatro etiquetas para cada examen (indicadas por yˆR,m, yˆL,m, yˆR,b y yˆL,b). Aunque estamos principalmente interesados en predecir con precisión la presencia o ausencia de hallazgos malignos, predecir la presencia o ausencia de hallazgos benignos cumple un papel importante de una tarea auxiliar que regulariza el aprendizaje de la tarea principal. Como entrada, tomamos cuatro imágenes de alta resolución correspondientes a las cuatro vistas de mamografía de detección estándar (indicadas por xR-CC, xL-CC, xR-MLO y xL-MLO). Recortamos cada imagen a un tamaño fijo de 2677 × 1942 píxeles para vistas CC y 2974 × 1748 píxeles para vistas MLO.3 Consulte la Fig. 3 para ver una representación esquemática.

IV. ARQUITECTURA MODELO Y ENTRENAMIENTO

Entrenamos CNN de múltiples vistas profundas de cuatro arquitecturas diferentes que se muestran en la Fig. 5, inspiradas en el trabajo anterior de Geras et al. [14]. Todas estas redes constan de dos módulos centrales: (i) cuatro columnas específicas de vista, cada una basada en la arquitectura ResNet [11] que genera una representación oculta de dimensión fija para cada vista de mamografía, y (ii) dos capas completamente conectadas para asigne las representaciones ocultas calculadas a las predicciones de salida. Los modelos difieren en cómo se agregan las representaciones ocultas específicas de la vista de todas las vistas para producir las predicciones finales. Consideramos las siguientes variantes.

1) El modelo de "vista inteligente" (Fig. 5(a)) concatena las representaciones L-CC y R-CC, y las representaciones L-MLO y R-MLO. Hace predicciones separadas para las vistas CC y MLO, que se promedian durante la inferencia.
2) El modelo 'image-wise' (Fig. 5(b)) hace una predicción para cada una de las cuatro vistas de forma independiente. Las predicciones correspondientes se promedian durante la inferencia.
3) El modelo 'lateral' (Fig. 5(c)) primero concatena las representaciones L-CC y L-MLO, y las representaciones R-CC y R-MLO, luego hace predicciones para cada seno por separado.
4) El modelo 'conjunto' (Fig. 5(d)) concatena las representaciones de las cuatro vistas y predice conjuntamente hallazgos malignos y benignos para ambas mamas.

En todos los modelos, utilizamos cuatro redes de 22 capas basadas en ResNet (ResNet-22) como columnas que calculan una red oculta de 256 dimensiones.

3Los tamaños y las ubicaciones de la ventana de recorte de cada imagen se ajustan para contener la mayor cantidad posible de tejido mamario mediante un método explicado en la sección 2.D del informe técnico sobre el conjunto de datos [13].

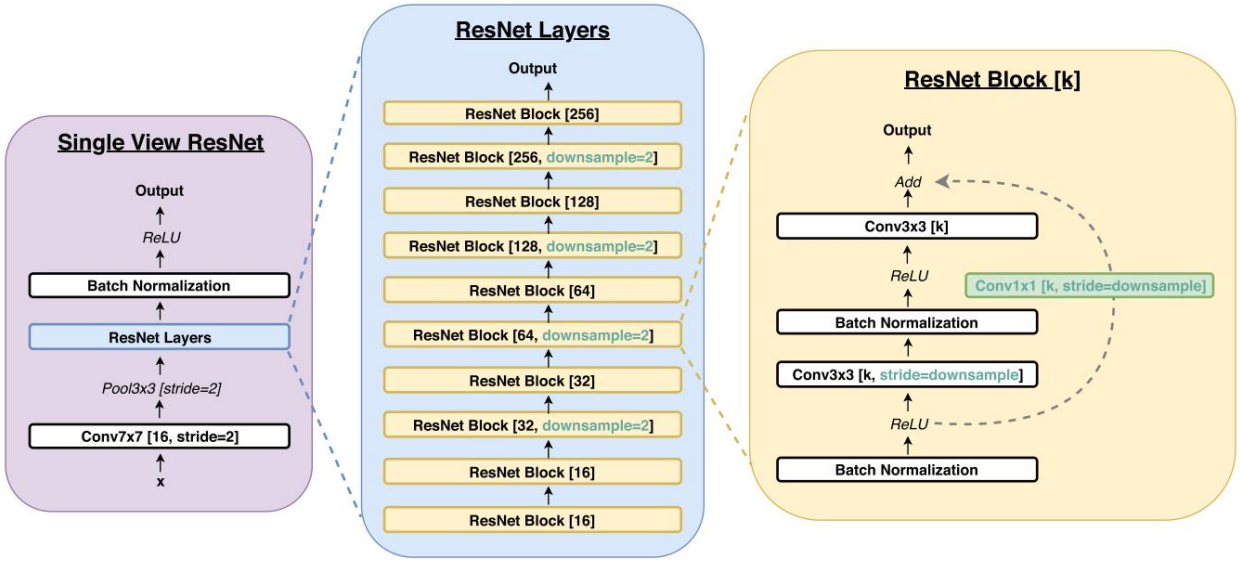

Fig. 4. Arquitectura de ResNet-22 de vista única. Los números entre corchetes indican el número de canales de salida, a menos que se especifique lo contrario. Izquierda: descripción general de ResNet-22 de vista única, que consta de un conjunto de capas de ResNet. Centro: las capas ResNet constan de una secuencia de bloques ResNet con diferentes canales de salida y reducción de resolución. Derecha: los bloques ResNet constan de dos capas convolucionales de 3 × 3, con operaciones de normalización por lotes y ReLU intercaladas, y una conexión residual entre la entrada y la salida. Cuando no se especifica un factor de reducción de muestreo para un bloque ResNet, la primera capa de convolución de 3 × 3 tiene un paso de 1 y se omite la operación de convolución de 1 × 1 para el residuo.

vector de representación de cada vista. En comparación con las ResNet estándar, esta red tiene un equilibrio diferente de profundidad y ancho, que se ajusta a imágenes de muy alta resolución.
Los detalles de la red ResNet-22 se encuentran en la Sección IV-A a continuación. Experimentalmente, encontramos que el modelo 'view-wise' es el más preciso en el conjunto de validación en términos de la tarea de predicción de malignidad/no malignidad. A menos que especifiquemos explícitamente lo contrario, informamos los resultados de este modelo.

A. ResNet-22 de vista única

La arquitectura completa de ResNet-22 se muestra en la Fig. 4. Vinculamos los pesos para L-CC y R-CC ResNets, así como para L-MLO y R-MLO ResNets.4 Del mismo modo, invertimos las imágenes L-CC y L-MLO antes de enviarlas al modelo, por lo que todas las imágenes de los senos están orientadas hacia la derecha, lo que permite que los pesos ResNet compartidos operen en imágenes orientadas de manera similar.

Una salida intermedia de cada ResNet es un tensor de 256 dimensiones H × W × donde H y W se reducen del tamaño de entrada original, con H = 42 y W = 31 para la vista CC, y H = 47 y W = 28 para vista MLO.

Promediamos esta representación a través de las dimensiones espaciales para obtener un vector de representación oculto de 256 dimensiones para cada vista. Como referencia, mostramos las dimensiones de las activaciones ocultas después de cada capa principal de ResNet-22 en la Tabla II.

La consideración principal al adaptar ResNets estándar para mamografías es la necesidad de procesar imágenes de muy alta resolución, sin reducción de resolución previa, ajustando el paso hacia adelante y el cálculo de gradiente dentro de la memoria GPU. Además, cada minilote procesado debe ser lo suficientemente grande para

4En la Sección IB del Material Complementario, mostramos resultados adicionales para un modelo de vista en el que los pesos para todas las vistas y lados están vinculados.

TABLA II

DIMENSIONES DE LOS MAPAS DE CARACTERÍSTICAS DESPUÉS DE CADA CAPA EN RESNET-22, SE MUESTRA COMO $H \times W \times D$. D INDICA EL NÚMERO DE FUNCIÓN MAPAS, H Y W INDICAR DIMENSIONES ESPACIALES

	CC view	MLO view
Conv7x7	1339×971×16	1487×874×16
ResBlock 1	670×486×16	744×437×16
ResBlock 2	335×243×32	372×219×32
ResBlock 3	168×122×64	186×110×64
ResBlock 4	84×61×128	93×55×128
ResBlock 5	42×31×256	47×28×256

el modelo de entrenamiento para estar bien acondicionado. Por ejemplo, descubrimos que la normalización de lotes afecta negativamente el entrenamiento para tamaños de minilotes menores a cuatro. Realizamos varios cambios para crear nuestro ResNet-22. Primero, debido a que las representaciones ocultas en las capas más bajas han sufrido la menor cantidad de reducción de resolución y, por lo tanto, son las de mayor tamaño, configuramos la primera capa convolucional para que tenga relativamente menos canales: 16 en comparación con 64 en los modelos estándar de Re
Para compensar, nuestro modelo tiene 5 bloques ResNet en comparación con los 4 de ResNet estándar. Como cada bloque ResNet duplica la cantidad de canales, nuestra representación oculta final tiene 256 canales, en comparación con los 512 en el caso de los modelos ResNet estándar. Efectivamente, aumentamos la capacidad entre canales más adelante en el modelo, intercambiando resoluciones más altas y menos canales al principio con representaciones ocultas más pequeñas y más canales más adelante en el modelo. Por último, mientras que en los modelos estándar de ResNet la capa de clasificación se aplica directamente después de la agrupación promedio global, en nuestro modelo, además, aplicamos dos capas completamente conectadas antes de la capa de clasificación. Hacemos esto para permitir interacciones más complejas entre diferentes vistas.

1) Entrenary Inferencia: Entrenamos todo el modelo usando el algoritmo de optimización de Adam [17], usando una tasa de aprendizaje de

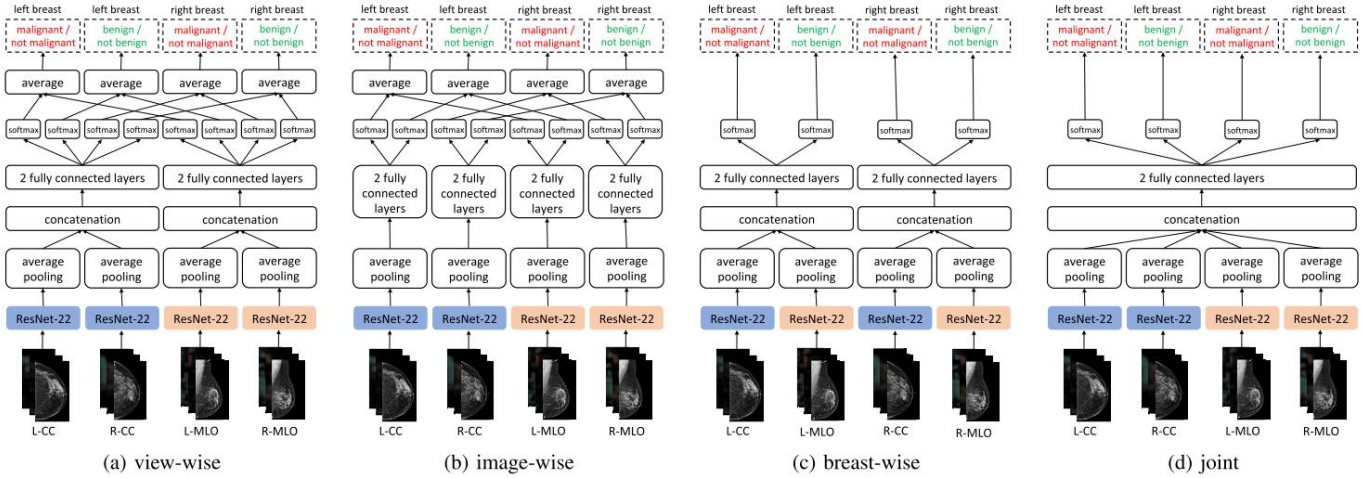

Fig. 5. Cuatro variantes de modelo para incorporar información a través de las cuatro vistas de mamografía de detección en un examen. Todas las variantes están limitadas a tener un total de 1024 activaciones ocultas entre capas totalmente conectadas. El modelo 'view-wise', que es el modelo principal utilizado en nuestros experimentos, contiene ramas de modelo separadas para vistas CC y MLO: promediamos las predicciones en ambas ramas. El modelo 'image-wise' tiene una rama de modelo para cada imagen, y de manera similar promediamos las predicciones. El modelo 'seno-sabio' tiene ramas separadas por seno (izquierda y derecha). El modelo 'conjunto' solo tiene una sola rama, que opera en las representaciones concatenadas de las cuatro imágenes. La agrupación promedio en todos los modelos está promediando globalmente a través de las dimensiones espaciales en todos los mapas de características. Cuando se agregan mapas de calor (cf. Sección IV-B) como canales adicionales a las entradas correspondientes, las primeras capas de las columnas se modifican en consecuencia.

10−5 y un mini lote de tamaño 4. Aplicamos la regularización L2 a los pesos de nuestro modelo con un coeficiente de 10−4.5. El modelo tiene 6.132.592 parámetros entrenables (6.135.728 al usar los mapas de calor como se describe en la Sección IV-B, la única diferencia entre ambas arquitecturas es el tamaño del kernel en la primera capa convolucional para acomodar la diferencia en el número de canales de entrada). En una GPU Nvidia V100, el modelo tarda aproximadamente 12 horas en entrenarse para obtener el mejor rendimiento de validación (24 horas cuando se usan los mapas de calor). Una cantidad significativa de sobrecarga de entrenamiento está asociada con el tiempo para cargar y aumentar las imágenes de mamografía de alta resolución. Los detalles sobre el aumento de datos se encuentran en la Sección III del Material complementario.

Solo una pequeña fracción de los exámenes en nuestro conjunto de entrenamiento contiene imágenes de senos biopsiados. El aprendizaje con datos muestreados uniformemente del conjunto de entrenamiento sería muy lento, ya que el modelo vería pocos ejemplos positivos por época. Para paliar este problema, dentro de cada época de entrenamiento, se mostraron en el modelo todos los exámenes con biopsias en el conjunto de entrenamiento (4844 exámenes), pero solo un subconjunto aleatorio de un número igual de exámenes sin biopsias (también 4844 exámenes). Detuvimos el entrenamiento antes de tiempo cuando el promedio de las AUC de validación en las cuatro tareas de predicción no mejoró durante 20 épocas. Luego, seleccionamos la versión del modelo con el mejor AUC de validación como nuestro modelo candidato final (mostramos la curva de entrenamiento y validación para un modelo de solo imagen y un modelo de imagen y mapas de calor en la Sección II-A en el Material complementario) .

En experimentos preliminares, notamos que al entrenar el modelo de vistas, optimizar la predicción para cada vista por separado conduce a una mejor generalización. Por lo tanto, aunque en el momento de la inferencia la predicción para cada seno se calcula como un promedio de las predicciones para ambas vistas de ese seno, el modelo en realidad se entrena para optimizar la pérdida, que trata las predicciones para las dos vistas por separado. Eso es, las predicciones para cada objetivo (como se define en la Sección III) se calculan como

$$\hat{y}_{R,m}(x_{R-CC}, x_{L-CC}, x_{R-MLO}, x_{L-MLO}) = \frac{1}{2}\hat{y}^{CC}_{R,m}(x_{R-CC}, x_{L-CC}) + \frac{1}{2}\hat{y}^{MLO}_{R,m}(x_{R-MLO}, x_{L-MLO}),$$

$$\hat{y}_{R,b}(x_{R-CC}, x_{L-CC}, x_{R-MLO}, x_{L-MLO}) = \frac{1}{2^{\text{año}}}\hat{y}^{CC}_{R,b}(x_{R-CC}, x_{L-CC}) + \frac{1}{2^{\text{año}}}\hat{y}^{MLO}_{R,b}(x_{R-MLO}, x_{L-MLO}),$$

$$\hat{y}_{L,m}(x_{R-CC}, x_{L-CC}, x_{R-MLO}, x_{L-MLO}) = \frac{1}{2}\hat{y}^{CC}_{L,m}(x_{R-CC}, x_{L-CC}) + \frac{1}{2}\hat{y}^{MLO}_{L,m}(x_{R-MLO}, x_{L-MLO}),$$

$$\hat{y}_{L,b}(x_{R-CC}, x_{L-CC}, x_{R-MLO}, x_{L-MLO}) = \frac{1}{2^{\text{año}}}\hat{y}^{CC}_{L,b}(x_{R-CC}, x_{L-CC}) + \frac{1}{2^{\text{año}}}\hat{y}^{MLO}_{L,b}(x_{R-MLO}, x_{L-MLO}),$$

mientras que la pérdida de entrenamiento se calcula como

$$\begin{aligned}
&L(y_{R,m}, y_{L,m}, y_{R,m}, y_{L,m}, x_{R-CC}, x_{L-CC}, x_{R-MLO}, x_{L-MLO}) \\
&= \ell(y_{R,m}, \hat{y}^{CC}_{R,m}(x_{R-CC}, x_{L-CC})) \\
&\quad + \ell(y_{R,m}, \hat{y}^{MLO}_{R,m}(x_{R-MLO}, x_{L-MLO})) \\
&\quad + \ell(y_{R,b}, \hat{y}^{CC}_{R,b}(x_{R-CC}, x_{L-CC})) \\
&\quad + \ell(y_{R,b}, \hat{y}^{MLO}_{R,b}(x_{R-MLO}, x_{L-MLO})) + \\
&\quad \ell(y_{L,m}, \hat{y}^{CC}_{L,m}(x_{R-CC}, x_{L-CC})) + \\
&\quad \ell(y_{L,m}, \hat{y}^{MLO}_{L,m}(x_{R-MLO}, x_{L-MLO})) \\
&\quad + \ell(y_{L,b}, \hat{y}^{CC}_{L,b}(x_{R-CC}, x_{L-CC})) \\
&\quad + \ell(y_{L,b}, \hat{y}^{MLO}_{L,b}(x_{R-MLO}, x_{L-MLO})),
\end{aligned}$$

donde ℓ denota entropía cruzada binaria.

La observación de que cuando una de las dos modalidades de entrada es más predictiva que la otra, la red tiende a ignorar la modalidad menos predictiva es consistente con los resultados anteriores [18]. En nuestros experimentos, encontramos que la vista CC

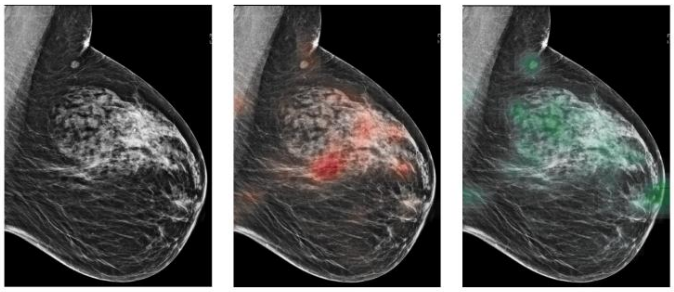

Fig. 6. La imagen original (izquierda), el mapa de calor 'maligno' sobre la imagen (centro) y el mapa de calor 'benigno' sobre la imagen (derecha).

es más predictivo que la vista MLO (consulte la Sección IC en el Material complementario).

B. Mapas de calor y modelo de clasificación de nivel de parche auxiliar

La alta resolución de las imágenes y la memoria limitada de las GPU nos obligan a usar ResNets relativamente superficiales dentro de nuestro modelo cuando usamos imágenes de resolución completa como entradas. Para aprovechar aún más los detalles de grano fino en las mamografías, entrenamos un modelo auxiliar para clasificar parches de mamografías de 256 × 256 píxeles, prediciendo la presencia o ausencia de hallazgos malignos y benignos en un parche determinado.
Las etiquetas de estos parches se determinan en función de las segmentaciones a nivel de píxel de las mamografías correspondientes producidas por los médicos. Nos referimos a este modelo como un modelo a nivel de parche , en contraste con el modelo a nivel de seno descrito en la sección anterior, que opera con imágenes de todo el seno.

Posteriormente, aplicamos esta red auxiliar a las mamografías de resolución completa en forma de ventana deslizante para crear dos mapas de calor para cada imagen (un ejemplo en la Fig. 6), uno que contiene una probabilidad estimada de un hallazgo maligno para cada píxel y el otro que contiene una probabilidad estimada de un hallazgo benigno. En total, obtenemos ocho imágenes adicionales: xm xm L−CC, xm R−MLO, $x_{b\,R\text{-}CC}$, $x_{b\,R\text{-}CC}$, $x_{b\,L\text{-}CC}$, $x_{b\,R\text{-}MLO}$, $x_{m\,L\text{-}MLO}$, $x_{b\,L\text{-}MLO}$. Estos parches clasifican Los mapas de calor de ción se pueden usar como canales de entrada adicionales para el modelo a nivel del seno para proporcionar información detallada adicional. Es decir, las entradas modificadas a la red entonces son: [xR−CC; xm R-CC; xb R−CC], [xL−CC; xm L-CC; xb L-CC], [xR-MLO; xm R-MLO; xb R−MLO], [xL−MLO; xm L-MLO; xb L−MLO].

El uso de modelos separados a nivel de píxel y de mama como se describe anteriormente diferencia nuestro trabajo de los enfoques que utilizan etiquetas a nivel de píxel en una única red diferenciable [19] o modelos basados en las variaciones de R-CNN [20]. Nuestro enfoque nos permite utilizar una red auxiliar muy profunda a nivel de parche, ya que esta red no tiene que procesar toda la imagen de alta resolución a la vez. Agregar los mapas de calor producidos por el clasificador de nivel de parche como canales de entrada adicionales permite que el clasificador principal obtenga el beneficio de las etiquetas de nivel de píxel, mientras que el cálculo pesado necesario para producir las predicciones de nivel de píxel no necesita repetirse cada vez que se muestra un ejemplo. utilizado para el aprendizaje. También podemos inicializar los pesos del clasificador a nivel de parche utilizando los pesos de las redes entrenadas previamente en grandes conjuntos de datos fuera del dominio, como

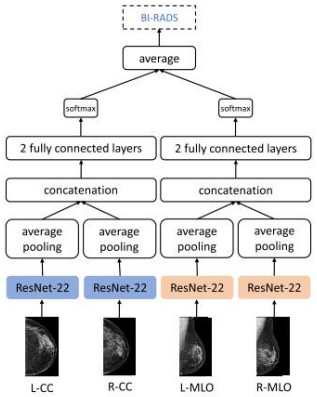

Fig. 7. Arquitectura del modelo de clasificación BI-RADS. La arquitectura es en gran medida similar a la variante del modelo de clasificación de cáncer "vista inteligente", excepto que la salida es un conjunto de estimaciones de probabilidad sobre las tres clases de salida. El modelo consta de cuatro columnas ResNet-22, con pesos compartidos dentro de las ramas CC y MLO del modelo.

ImageNet [21].5 De aquí en adelante, nos referiremos al modelo que usa solo imágenes de mamografía como modelo de solo imagen , y al modelo que usa imágenes de mamografía y mapas de calor como modelo de imagen y mapas de calor .

C. Pre-entrenamiento Clasificación BI-RADS en

Debido a la cantidad relativamente pequeña de ejemplos de biopsias con etiquetas benignas o malignas que tenemos disponibles, aplicamos el aprendizaje por transferencia para mejorar la solidez y el rendimiento de nuestros modelos. Transferir el aprendizaje implica reutilizar partes de un modelo entrenado previamente en otra tarea como punto de partida para entrenar el modelo de destino, aprovechando las representaciones aprendidas de la tarea de entrenamiento previo.

Para nuestro modelo, aplicamos el aprendizaje de transferencia de una red previamente entrenada en una tarea de clasificación BI-RADS, como en [14], que corresponde a predecir la evaluación de un radiólogo del riesgo de un paciente de tener cáncer de mama basándose únicamente en la mamografía de detección. Las tres clases de BI-RADS que consideramos son: Categoría 0 de BI-RADS ("incompleta"), Categoría 1 de BI-RADS ("normal") y Categoría 2 de BI-RADS ("benigna").
El algoritmo utilizado para extraer estas etiquetas se explica en [13]. Aunque estas etiquetas son más ruidosas que los resultados de la biopsia (ya que las evaluaciones de los médicos se basan en mamografías de detección y no se basan en una biopsia), en comparación con los 4844 exámenes con etiquetas de cáncer comprobadas por biopsia en el conjunto de capacitación, tenemos más de 99 528 ejemplos de capacitación con Etiquetas BI RADS 0 y BI-RADS 2. Las redes neuronales han sido demostrado alcanzar niveles razonables de rendimiento incluso cuando se entrenan con etiquetas ruidosas [22], [23]. Usamos esta propiedad para transferir la información aprendida con las etiquetas BI-RADS al modelo de clasificación del cáncer. De hecho, nuestros experimentos muestran que el entrenamiento previo en la clasificación BI-RADS contribuye significativamente al desempeño de nuestro modelo (ver Sección VE).

El modelo que utilizamos para la clasificación BI-RADS se muestra en la Fig. 7. Es similar a la arquitectura del modelo "vista inteligente" para la clasificación del cáncer descrita en la sección Variantes del modelo .

5Para ajustar una red previamente entrenada en imágenes RGB con imágenes en escala de grises, duplicamos las imágenes en escala de grises a través de los canales RGB.

TABLA III

AUCS DE NUESTROS MODELOS SOBRE POBLACIONES TAMIZADAS Y BIOPSIADAS . TODOS LOS MODELOS, EXCEPTO LOS INDICADOS CON * FUERON PREENTRENADOS EN LA CLASIFICACIÓN BI-RADS

		screening population				**biopsied population**			
		single		5x ensemble		single		5x ensemble	
		malignant	benign	malignant	benign	malignant	benign	malignant	benign
image-only	view-wise	0.827±0.008	0.731±0.004	0.840	0.743	0.781±0.006	0.673±0.003	0.791	0.682
	view-wise*	0.687±0.009	0.657±0.006	0.703	0.669	0.693±0.006	0.564±0.006	0.709	0.571
	image-wise	0.830±0.006	0.759±0.002	0.841	0.766	0.740±0.007	0.638±0.001	0.749	0.642
	breast-wise	0.821±0.012	0.757±0.002	0.836	0.768	0.726±0.009	0.639±0.002	0.738	0.645
	joint	0.822±0.008	0.737±0.004	0.831	0.746	0.780±0.006	0.682±0.001	0.787	0.688
image-and-heatmaps	view-wise	**0.886±0.003**	0.747±0.002	**0.895**	0.756	**0.843±0.004**	0.690±0.002	**0.850**	0.696
	view-wise*	0.856±0.007	0.701±0.004	0.868	0.708	0.828±0.008	0.633±0.006	0.841	0.640
	image-wise	0.875±0.001	**0.765±0.003**	0.885	0.774	0.812±0.001	0.653±0.003	0.821	0.658
	breast-wise	0.876±0.004	0.764±0.004	0.889	**0.779**	0.805±0.004	0.652±0.004	0.818	0.661
	joint	0.860±0.008	0.745±0.002	0.876	0.763	0.817±0.008	**0.696±0.005**	0.830	**0.709**

anterior, excepto que la capa de salida genera estimaciones de probabilidad sobre tres clases para una sola etiqueta. Medimos el rendimiento de este modelo promediando las AUC de 0 frente a otras predicciones, 1 frente a otras y 2 frente a otras predicciones en el conjunto de validación.

El resto de los detalles del entrenamiento (p. ej., arquitectura ResNet-22, hiperparámetros del optimizador) son idénticos a los del modelo de clasificación del cáncer, excepto que el modelo se entrenó con un tamaño de minilote de 24 en lugar de 4. Detuvimos el entrenamiento antes de tiempo según AUC de validación después de no mejorar durante 20 épocas, e inicializó los pesos ResNet-22 para el modelo de clasificación del cáncer utilizando los pesos aprendidos en el modelo BI-RADS. Donde usamos mapas de calor como canales de entrada adicionales, duplicamos los pesos en el núcleo convolucional más bajo de modo que el modelo pueda operar en entradas con tres canales; el resto del modelo permanece sin cambios. En nuestros resultados experimentales, utilizamos un modelo BI-RADS entrenado para 111 épocas (326 horas en cuatro GPU Nvidia V100), que obtuvo un AUC de validación promedio de 0,748.

Enfatizamos aquí que usamos las mismas divisiones de prueba de validación de trenes para el entrenamiento previo de nuestro modelo de clasificación BI-RADS como en el entrenamiento de nuestro modelo de clasificación de cáncer, por lo que no fue posible la fuga de datos entre las divisiones.

V. EXPERIMENTOS

En todos los experimentos, usamos el conjunto de entrenamiento para optimizar los parámetros de nuestro modelo y el conjunto de validación para ajustar los hiperparámetros del modelo y el procedimiento de entrenamiento. A menos que se especifique lo contrario, los resultados se calcularon en toda la población de cribado. Para obtener predicciones para cada ejemplo de prueba, aplicamos transformaciones aleatorias a la entrada 10 veces, aplicamos el modelo a cada una de las 10 muestras por separado y luego promediamos las 10 predicciones (detalles en la Sección III del Material complementario).

Para mejorar aún más nuestros resultados, empleamos la técnica de ensamblaje de modelos [24], en la que se promedian las predicciones de varios modelos diferentes para producir la predicción general del conjunto. En nuestro caso, entrenamos cinco copias de cada modelo con diferentes inicializaciones aleatorias de los pesos en las capas totalmente conectadas, mientras que los pesos restantes se inicializan con los pesos del modelo preentrenado en la clasificación BI-RADS. Para cada modelo, reportamos los resultados. de una sola red (media y desviación estándar a través de cinco inicializaciones aleatorias) y de un conjunto.

A. Poblaciones de prueba

En los experimentos a continuación, evaluamos nuestro modelo en varias poblaciones para probar diferentes hipótesis: (i) población de detección, incluidos todos los exámenes del conjunto de prueba sin submuestreo; (ii) subpoblación sometida a biopsia, que es un subconjunto de la población de cribado, que solo incluye exámenes de la población de cribado que contiene mamas que se sometieron a una biopsia; (iii) subpoblación de estudio de lectores, que consiste en la subpoblación de la biopsia y un subconjunto de exámenes muestreados aleatoriamente de la población de detección sin ningún hallazgo.

B. Métricas de evaluación

Evaluamos nuestros modelos principalmente en términos de AUC (área bajo la curva ROC) para las tareas de clasificación de maligno/no maligno y benigno/no benigno a nivel mamario. El modelo y las respuestas de los lectores en el subconjunto para el estudio del lector se evalúan en términos de AUC, así como AUC de recuperación de precisión (PRAUC), que son métricas comúnmente utilizadas en la evaluación del desempeño de los radiólogos. ROC y PRAUC capturan diferentes aspectos del rendimiento de un modelo predictivo.

La curva ROC resume el equilibrio entre la tasa de verdaderos positivos y la tasa de falsos positivos para un modelo que utiliza diferentes umbrales de probabilidad. La curva de recuperación de precisión resume el equilibrio entre la tasa positiva verdadera (recuperación) y el valor predictivo positivo (precisión) para un modelo que usa diferentes umbrales de probabilidad.

C. Población examinada

En este apartado presentamos los resultados sobre la población cribada, que se aproxima a la distribución de los pacientes que se someten a cribado rutinario. Los resultados de las diferentes variantes del modelo se muestran en la Tabla III. En general, las cuatro variantes del modelo logran AUC altas y relativamente similares. El conjunto de mapas de calor e imágenes de 'vista inteligente', que también es arquitectónicamente más similar al modelo BI-RADS utilizado en la etapa previa al entrenamiento, se desempeña mejor en la predicción de malignidad/no malignidad,

alcanzando un AUC de 0,895 en la población cribada y de 0,850 en la población biopsiada. Sin embargo, algunas de las otras variantes del modelo superan el conjunto 'visto-sabio' para la predicción benigna/no benigna. Entre los modelos de solo imagen, las cuatro variantes del modelo tienen un rendimiento más o menos comparable, aunque siguen teniendo un rendimiento inferior al de los modelos de mapas de calor y de imagen. Los modelos de imágenes y mapas de calor mejoran más en la clasificación maligno/no maligno que en la clasificación benigno/no benigno. También encontramos que el ensamblaje es beneficioso en todos los modelos, lo que lleva a un aumento pequeño pero constante en AUC.

La construcción de un conjunto de las cuatro variantes del modelo para el modelo de imágenes y mapas de calor, con cinco modelos inicializados aleatoriamente por variante, da como resultado un AUC de 0,778 en la predicción benigna/no benigna y de 0,899 en la predicción maligna/no maligna en el cribado población. Aunque este rendimiento es superior a cualquier variante de modelo individual, ejecutar un conjunto tan grande de 20 modelos separados sería prohibitivamente costoso en la práctica.

La discrepancia en el desempeño de nuestros modelos entre las tareas malignas/no malignas y benignas/no benignas puede explicarse en gran medida por el hecho de que una fracción mayor de hallazgos benignos que de hallazgos malignos están mamográficamente ocultos (Tabla I) . Además, puede haber ruido en las etiquetas benigno/no benigno asociado con la confianza de los radiólogos en sus diagnósticos. Para el mismo examen, un radiólogo podría descartar un hallazgo como obviamente no maligno sin solicitar una biopsia, mientras que otro radiólogo podría ser más conservador y solicitar una biopsia.

Usando el conjunto de validación, encontramos que el modelo de mapas de calor e imágenes 'vistas' supera a todas las demás variantes en términos del promedio de AUC para tareas de predicción malignas/no malignas y benignas/no benignas. A menos que se especifique lo contrario, tanto para el modelo de solo imagen como para el de imagen y mapas de calor, nos referimos a los resultados basados en el modelo de 'vista' en las siguientes secciones.

D. Subpoblación biopsiada

Mostramos los resultados de nuestros modelos evaluados solo en la subpoblación biopsiada, en la mitad derecha de la Tabla III. Dentro de nuestro conjunto de pruebas, esto corresponde a 401 senos: 339 con hallazgos benignos, 45 con hallazgos malignos y 17 con ambos. Esta subpoblación que se sometió a una biopsia con al menos un hallazgo de imágenes difiere notablemente de la población general de exámenes de detección, que consiste en personas en gran parte sanas que se someten a exámenes de detección anuales de rutina sin necesidad de exámenes de diagnóstico por imágenes o biopsias adicionales. En comparación con los resultados de la población de cribado, las AUC de la población sometida a biopsia son notablemente más bajas en todas las variantes del modelo.

En la subpoblación biopsiada, observamos una diferencia constante entre el rendimiento de los modelos de solo imagen e imagen y mapas de calor. El conjunto de modelos de imágenes y mapas de calor funciona mejor tanto en la clasificación de maligno/no maligno, con un AUC de 0,850, como en la clasificación de benigno/no benigno, con un AUC de 0,696. Las AUC marcadamente más bajas alcanzadas para la subpoblación biopsiada, en comparación con la población de cribado, pueden explicarse por el hecho de que los exámenes que requieren un retiro para el diagnóstico por imágenes y que posteriormente necesitan una biopsia son más desafiantes tanto para los radiólogos como para nuestro modelo.6

MI. Importancia de Pre-entrenamiento Clasificación BI-RADS en

En esta sección, evaluamos el beneficio del entrenamiento previo de BI-RADS al comparar el rendimiento de nuestros modelos con los modelos de clasificación de cáncer entrenados sin usar pesos de un modelo BI-RADS previamente entrenado. Específicamente, entrenamos un conjunto de modelos de clasificación de cáncer a partir de pesos de modelo inicializados completamente al azar.

Los resultados se muestran en la Tabla III (marcados con *). En todos los casos, vemos una mejora en el rendimiento al usar los pesos de un modelo entrenado previamente en la clasificación BI-RAD, en comparación con la inicialización aleatoria de los pesos del modelo y el entrenamiento desde cero. La mejora en el rendimiento del uso de pesos previamente entrenados tiende a ser mayor para el modelo de solo imagen en comparación con los modelos de imagen y mapas de calor. Nuestra hipótesis es que esto se debe a que los mapas de calor ya contienen información importante relacionada con la clasificación del cáncer y, por lo tanto, es probable que el modelo aprenda más rápidamente a utilizar los mapas de calor para la clasificación del cáncer. Por el contrario, los modelos de solo imagen se basan completamente en ResNets para codificar de manera efectiva la información visual para la clasificación del cáncer y, por lo tanto, el uso de los pesos de un modelo entrenado previamente para la clasificación BI-RADS contribuye significativamente al rendimiento del modelo.

VI. ESTUDIO DEL LECTOR

Para comparar el rendimiento de nuestro conjunto de imágenes y mapas de calor (en lo sucesivo, el modelo) con radiólogos humanos, realizamos un estudio de lectores con 14 lectores: 12 radiólogos asistentes con varios niveles de experiencia (entre 2 y 25 años), un residente y un estudiante de medicina, cada uno de los cuales leyó 740 exámenes del conjunto de pruebas (1480 senos): 368 exámenes seleccionados al azar de la subpoblación biopsiada y 372 exámenes seleccionados al azar de exámenes que no coincidían con ninguna biopsia. Los exámenes se barajaron antes de entregarlos a los lectores. Se pidió a los lectores que proporcionaran una estimación de la probabilidad de malignidad en una escala del 0 % al 100 % para cada seno.
Como algunos senos contienen múltiples hallazgos sospechosos, se pidió a los lectores que dieran su evaluación del hallazgo más sospechoso.

Usamos los primeros 20 exámenes como un conjunto de prácticas para familiarizar a los lectores con el formato del estudio del lector; estos fueron excluidos del análisis.7 En los 720 exámenes restantes, evaluamos el desempeño del modelo y de los lectores en la clasificación de malignidad. Entre los 1.440 pechos, hay

6Más precisamente, esta diferencia en el AUC puede explicarse por el hecho de que si bien agregar o restar ejemplos negativos a la población de prueba no cambia la tasa de verdaderos positivos, sí altera la tasa de falsos positivos. La tasa de falsos positivos se calcula como una proporción de falsos positivos y negativos. Por lo tanto, al agregar ejemplos negativos fáciles al conjunto de pruebas, la cantidad de falsos positivos crecerá más lentamente que la cantidad de todos los negativos, lo que conducirá a un aumento en el AUC. Por otro lado, eliminar ejemplos negativos fáciles tendrá un efecto inverso y el AUC será menor.

7A los lectores se les mostraron las imágenes y se les pidió que dieran su valoración. Confirmamos la corrección del formato en el que devolvieron sus respuestas, pero no les proporcionamos comentarios sobre la precisión de sus predicciones.

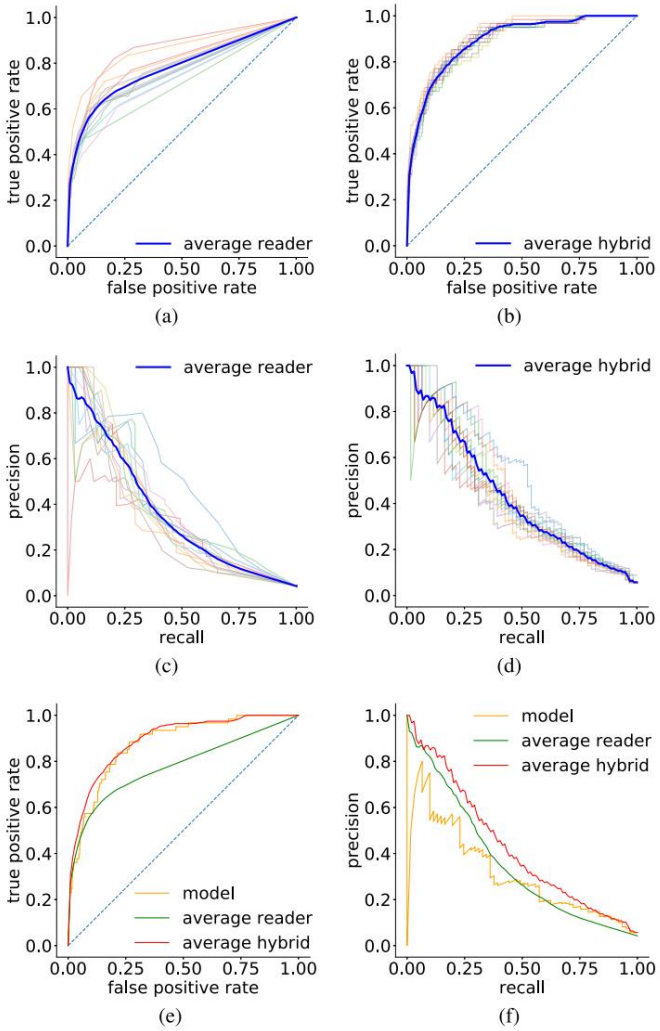

Fig. 8. Curvas ROC [(a), (b) y (e)] y curvas de recuperación de precisión [(c), (d) y (f)] en el subconjunto del conjunto de prueba utilizado para el lector estudiar. Curvas (a) y (c) para los 14 lectores. Su desempeño promedio está resaltado en azul. (b) y (d) curvas para el híbrido del conjunto de imágenes y mapas de calor con cada lector individual. La curva resaltada en azul indica el rendimiento promedio de todos los híbridos. (e) y (f) comparación entre el conjunto de imágenes y mapas de calor, el lector promedio y el híbrido promedio.

62 mamas etiquetadas como malignas y 356 mamas etiquetadas como benignas. En las mamas etiquetadas como malignas hay 21 masas, 26 calcificaciones, 12 asimetrías y 4 distorsiones arquitectónicas89. En las mamas etiquetadas como benignas los números correspondientes de hallazgos por imagen son: 87, 102, 36 y 6.

Nuestro modelo logró un AUC de 0,876 y PRAUC de 0,318.

Las AUC logradas por lectores individuales variaron de 0,705 a 0,860 (media: 0,778, estándar: 0,0435). PRAUCs para lectores variados

8 Las masas se definen como lesiones tridimensionales que ocupan un espacio con bordes completa o parcialmente convexos hacia afuera. Las calcificaciones son pequeñas motas de depósitos calcificados. Una asimetría se define como un depósito unilateral de tejido fibroglandular que no cumple con la definición de masa, es decir, es un área del tejido fibroglandular que no se ve en el otro seno. La distorsión arquitectónica se refiere a una interrupción del patrón aleatorio normal del tejido fibroglandular sin una masa definida visible.

9Como un seno tenía dos tipos de hallazgos, los números suman 63, no 62.

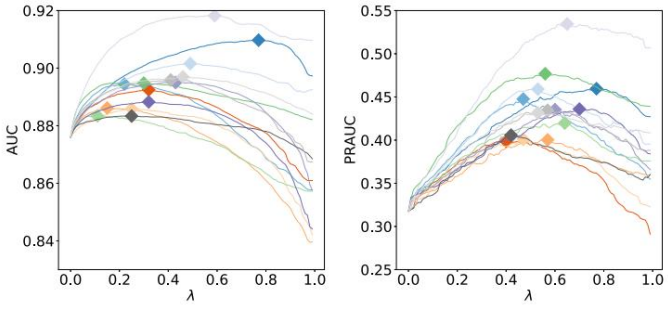

Fig. 9. AUC (izquierda) y PRAUC (derecha) en función de λ [0, 1) para híbridos entre cada lector y nuestro conjunto de imágenes y mapas de calor. Cada híbrido logra el AUC/PRAUC más alto para un λ diferente (marcado con ♦).

de 0,244 a 0,453 (media: 0,364, estándar: 0,0496). Las ROC individuales y las curvas de recuperación de precisión, junto con sus promedios, se muestran en la Fig. 8(a) y la Fig. 8(c).

También evaluamos la precisión de un híbrido hombre-máquina, cuyas predicciones son una combinación lineal de las predicciones de un radiólogo y del modelo, es decir,

$$\hat{y}_{\text{híbrido}} = \lambda\ \hat{y}_{\text{radiólogo}} + (1 - \lambda\ )\hat{y}_{\text{modelo}}.$$

Para λ = 0,510 (ver Fig. 9 para los resultados de λ [0, 1)), los híbridos entre cada lector y el modelo lograron un AUC promedio de 0,891 (estándar: 0,0109) y un PRAUC promedio de 0,431 (estándar: 0,0332) (cf. Fig. 8(b), Fig. 8(d)). Estos resultados sugieren que nuestro modelo puede usarse como una herramienta para ayudar a los radiólogos a leer los exámenes de detección del cáncer de mama y que capturó diferentes aspectos de la tarea en comparación con los radiólogos de mama experimentados. Un análisis cualitativo que compara las predicciones hechas por nuestra red y por los radiólogos para exámenes específicos se puede encontrar en la Sección IG-1 en el Material con

A. Visualización de la Representación Aprendida por el Clasificador

Además, examinamos cómo la red representa los exámenes internamente al visualizar las representaciones ocultas aprendidas por el mejor modelo de imagen única y mapas de calor, para exámenes en la subpoblación de estudio de lectores. Visualizamos dos conjuntos de activaciones: activaciones concatenadas de la última capa de cada una de las cuatro columnas específicas de la imagen y activaciones concatenadas de la primera capa completamente conectada en las ramas del modelo CC y MLO. Ambos conjuntos de activaciones tienen 1.024 dimensiones en total. Los incrustamos en un espacio bidimensional usando UMAP [25] con la distancia euclidiana.

La Fig. 10 muestra los puntos incrustados. El color y el tamaño de cada punto reflejan la misma información: cuanto más cálido y grande es el punto, mayor es la predicción media de malignidad de los lectores. Se calcula una puntuación para cada examen como un promedio sobre las predicciones para los dos senos. observamos que

10No tenemos una forma de ajustar λ a lectores individuales, por lo tanto, elegimos λ = 0,5 como la forma más natural de agregar dos conjuntos de predicciones cuando no se tiene conocimiento previo de su calidad. Como muestra la Fig. 9 , un λ óptimo varía mucho según el lector. Cuanto más fuerte sea el rendimiento del lector, menor será el peso óptimo en el modelo. Cabe destacar que todos los lectores pueden mejorar al promediar sus predicciones con el modelo para ambas métricas

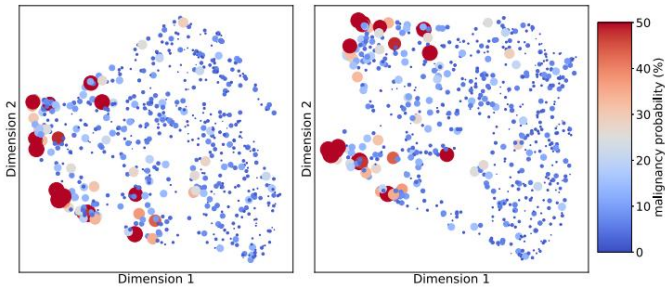

Fig. 10. Proyección UMAP bidimensional de las activaciones computadas por la red para los exámenes en el estudio de lectores. Visualizamos dos conjuntos de activaciones: (izquierda) activaciones concatenadas de la última capa de cada una de las cuatro columnas específicas de la imagen y (derecha) activaciones concatenadas de la primera capa completamente conectada en las ramas del modelo CC y MLO. Cada punto representa un examen. El color y el tamaño de cada punto reflejan la misma información: probabilidad de malignidad predicha por los lectores (promedio de los dos senos y los 14 lectores).

los exámenes clasificados como más propensos a ser malignos según los lectores están cerca uno del otro para ambos conjuntos de activaciones.

El hecho de que la red descubriera que los exámenes de neoplasias malignas nunca antes vistos eran similares corrobora aún más que nuestro modelo exhibe fuertes capacidades de generalización.

VIII. TRABAJO RELACIONADO

Trabajos previos abordan la tarea de clasificación de exámenes de detección de cáncer de mama en dos paradigmas. En un paradigma, solo están disponibles las etiquetas a nivel de examen, a nivel de mama o a nivel de imagen. Primero se aplica una CNN a cada una de las cuatro vistas estándar y los vectores de características resultantes se combinan para producir una predicción final [14]. Este flujo de trabajo se puede integrar aún más con el aprendizaje multitarea donde se pueden incorporar evaluaciones radiológicas, como la densidad mamaria, para modelar la confianza de la clasificación [26]. Otros trabajos formulan la tarea de clasificación del examen de cáncer de mama como una localización débilmente supervisada y producen un mapa de activación de clases que destaca las ubicaciones de las lesiones sospechosas [27]. Dichas formulaciones se pueden combinar con el aprendizaje de instancias múltiples en el que cada ubicación espacial se trata como una instancia única y se asocia con una puntuación que se correlaciona con la existencia de un hallazgo maligno [28].

En el segundo paradigma, las etiquetas a nivel de píxel que indican la ubicación de hallazgos benignos o malignos también se proporcionan al clasificador durante el entrenamiento. Las etiquetas a nivel de píxel permiten entrenar modelos derivados de la arquitectura R-CNN [20] o modelos que dividen las mamografías en parches más pequeños y entrenan clasificadores a nivel de parche usando la ubicación de hallazgos malignos [19], [29]–[32] . Algunos de estos trabajos agregan directamente las salidas del clasificador de nivel de parche para formar una predicción de nivel de imagen. Una limitación importante de tales arquitecturas es que se descuidará la información fuera de las regiones de interés anotadas. Otros trabajos aplican el clasificador a nivel de parche como un primer nivel de extracción de características sobre el cual se apilan más capas y luego se optimiza todo el modelo de forma conjunta.
Una desventaja de este tipo de arquitectura es el requisito de que todo el modelo quepa en la memoria de la GPU para el entrenamiento, lo que limita el tamaño del minilote utilizado (normalmente a uno), la profundidad del modelo a nivel de parche y la densidad del nivel de parche. se aplica el modelo. Nuestro trabajo es más similar al último tipo de modelos que utilizan etiquetas a nivel de píxel, sin embargo, nuestra estrategia utiliza un clasificador a nivel de parche para producir mapas de calor como canales de entrada adicionales para el clasificador a nivel de mama. Si bien renunciamos a la capacidad de entrenar todo el modelo de un extremo a otro, el clasificador de nivel de parche puede ser significativamente más poderoso y puede aplicarse densamente en la imagen original. Como resultado, nuestro modelo tiene la capacidad de aprender tanto las características locales en toda la imagen como las características macroscópicas, como la simetría entre los senos. Para una revisión más completa del trabajo anterior, consulte una de las revisiones recientes [33], [34].

Se han informado una variedad de resultados en términos de AUC para la predicción de malignidad. Los más comparables a nuestro trabajo son: [28] (0,86), [20] (0,95), [35] (0,81), [27] (0,91), [36] (0,84) y [37] (0,89). Desafortunadamente, aunque estos resultados pueden servir como una estimación aproximada de la calidad del modelo, sería engañoso comparar diferentes métodos basados en estos números. Algunos autores no discuten el diseño de sus modelos [35]–[37], algunos evalúan sus modelos en conjuntos de datos públicos muy pequeños, InBreast [38] o DDSM [39], que son insuficientes para una evaluación significativa, mientras que otros utilizan datos privados. conjuntos de datos con poblaciones de diferentes distribuciones (en un espectro entre la población de cribado y la subpoblación biopsiada), diferentes calidades de equipos de imágenes e incluso etiquetas definidas de manera diferente. Al hacer públicos el código y los pesos de nuestro modelo, buscamos permitir comparaciones más directas con nuestro trabajo.

VIII. DISCUSIÓN Y CONCLUSIONES

Al aprovechar un gran conjunto de entrenamiento con etiquetas a nivel de mama y de píxel, creamos una red neuronal que puede clasificar con precisión los exámenes de detección del cáncer de mama. Atribuimos este éxito a la cantidad significativa de computación encapsulada en el modelo de nivel de parche, que se aplicó densamente a las imágenes de entrada para formar mapas de calor como canales de entrada adicionales para un modelo de nivel de pecho. Sería imposible entrenar este modelo de manera completamente integral con el hardware disponible actualmente. Aunque nuestros resultados son prometedores, reconocemos que el conjunto de pruebas utilizado en nuestros experimentos es relativamente pequeño y nuestros resultados requieren una mayor validación clínica. También reconocemos que aunque el desempeño de nuestra red es más fuerte que el de los radiólogos en la tarea específica de nuestro estudio de lectores, esta no es exactamente la tarea que realizan los radiólogos. Por lo general, la mamografía de detección es solo el primer paso en una tubería de diagnóstico, y el radiólogo toma una determinación final y toma la decisión de realizar una biopsia solo después de la revisión para obtener imágenes de mamografía de diagnóstico adicionales y una posible ecografía. Sin embargo, en nuestro estudio, un modelo híbrido que incluye tanto una red neuronal como radiólogos expertos se realizó individualmente, lo que sugiere que el uso de dicho modelo podría mejorar la sensibilidad del radiólogo para la detección del cáncer de mama.

Por otro lado, el diseño de nuestro modelo es relativamente simple. Son posibles modelos más sofisticados y precisos.
Además, la tarea que consideramos en este trabajo, predecir si el paciente tenía un cáncer visible en el momento de la mamografía de detección, es la más simple posible entre muchas tareas de interés. Además de probar la utilidad de

este modelo en la lectura en tiempo real de las mamografías de detección, un próximo paso claro sería predecir el desarrollo de cáncer de mama en el futuro, incluso antes de que sea visible para un ojo humano entrenado.

RECONOCIMIENTO

Los autores desean agradecer a Catriona C. Geras por corregir las versiones anteriores de este manuscrito, a Michael Cantor por proporcionarles informes de patología, a Marc Parente y Eli Bogom-Shanon por su ayuda con la importación de datos de imágenes y a Mario Videna por respaldar nuestro entorno informático. También agradecen el apoyo de Nvidia Corporation con la donación de algunas de las GPU utilizadas en esta investigación.

REFERENCIAS

[1] RL Siegel, KD Miller y A. Jemal, "Estadísticas del cáncer, 2015", CA, Cancer J. Clinicians, vol. 65, núm. 1, págs. 5 a 29, 2015.

[2] SW Duffy et al., "El impacto de la detección del servicio de mamografía organizado en la mortalidad por carcinoma de mama en siete condados suecos: una evaluación colaborativa", Cancer. vol. 95, núm. 3, págs. 458 y 469, 2002.

[3] DB Kopans, "Más allá de los ensayos controlados aleatorios: el cribado mamográfico organizado reduce sustancialmente la mortalidad por carcinoma de mama" Cáncer, vol. 94, núm. 2, págs. 580 y 581, 2002.

[4] SW Duffy, L. Tabár y RA Smith, "Los ensayos de detección mamográfica: comentario sobre el trabajo reciente de Olsen y Gøtzsche", CA, Cancer J. Clinicians, vol. 52, núm. 2, págs. 68–71, 2002.

[5] DB Kopans, "Una carta abierta a los paneles que deciden las pautas para la detección del cáncer de mama", Breast Cancer Res Treat, vol. 151, núm. 1, págs. 19 a 25, 2015.

[6] CD Lehman, RD Wellman, DSM Buist, K. Kerlikowske, ANA Tosteson y DL Miglioretti, "Exactitud diagnóstica de la mamografía de detección digital con y sin detección asistida por computadora" JAMA Internal Med., vol. 175, núm. 11, págs. 1828–1837, 2015.

[7] Y. LeCun, Y. Bengio y G. Hinton, "Aprendizaje profundo", Nature, vol. 521, págs. 436–444, mayo de 2015.

[8] Y. LeCun et al., "Reconocimiento de dígitos escritos a mano con una red de retropropagación", en Proc. NIPS, 1989, págs. 396–404.

[9] A. Krizhevsky, I. Sutskever y GE Hinton, "Clasificación de ImageNet con redes neuronales convolucionales profundas", en Proc. NIPS, 2012, págs. 1097–1105.

[10] K. Simonyan y A. Zisserman, "Redes convolucionales muy profundas para el reconocimiento de imágenes a gran escala", en Proc. ICLR, 2014, págs. 1–14.

[11] K. He, X. Zhang, S. Ren y J. Sun, "Aprendizaje residual profundo para el reconocimiento de imágenes", en Proc. CVPR, junio de 2016, págs. 770–778.

[12] G. Huang, Z. Liu, L. van der Maaten y KQ Weinberger, "Redes convolucionales densamente conectadas", en Proc. CVPR, junio de 2017, págs. 4700–4708.

[13] N. Wu et al., "El conjunto de datos de detección de cáncer de mama de la NYU V1.0", Universidad de Nueva York, Nueva York, NY, EE. UU., Tech. Rep., 2019. [En línea]. Disponible: https://cs.nyu.edu/~kgeras/reports/datav1.0.pdf

[14] KJ Geras et al., "Detección de cáncer de mama de alta resolución con redes neuronales convolucionales profundas multivista", 2017, arXiv :1703.07047. [En línea]. Disponible: https://arxiv.org/abs/1703.07047 [15] N. Wu et al., "Clasificación de densidad mamaria con redes neuronales convolucionales profundas", en Proc. ICASSP, abril de 2018, págs. 6682–6686.

[16] R. Caruana, "Aprendizaje multitarea", Mach. Aprende., vol. 28, núm. 1, págs. 41 a 75, 1997.

[17] DP Kingma y J. Ba, "Adam: Un método para la optimización estocástica," en Proc. ICLR, 2015, págs. 1–15.

[18] F. Li, N. Neverova, C. Wolf y G. Taylor, "Modout: aprendizaje de arquitecturas multimodales mediante regularización estocástica", en Proc. 12º IEEE internacional. Conf. automático Reconocimiento de gestos faciales, may/jun. 2017, págs. 422–429.

[19] W. Lotter, G. Sorensen y D. Cox, "Una estrategia de aprendizaje de currículo y CNN de múltiples escalas para la clasificación de mamografías", en Proc. DLMIA, 2017, págs. 169–177.

[20] D. Ribli, A. Horváth, Z. Unger, P. Pollner e I. Csabai, "Detectar y clasificar lesiones en mamografías con aprendizaje profundo", Sci. Rep., vol. 8 de marzo de 2018, art. No. 4165.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li y L. Fei-Fei, "ImageNet: una base de datos de imágenes jerárquicas a gran escala", en Proc. CVPR, 2009, págs. 1–8.

[22] J. Krause et al., "La efectividad irrazonable de los datos ruidosos para el reconocimiento detallado", 2015, arXiv: 1511.06789. [En línea]. Disponible: https://arxiv.org/abs/1511.06789 [23] C. Sun, A. Shrivastava, S. Singh y A. Gupta, "Revisando la efectividad irrazonable de los datos en la era del aprendizaje profundo", 2017, arXiv:1707.02968 . [En línea]. Disponible: https://arxiv.org/abs/1707.02968 [24] TG Dietterich, "Métodos de conjunto en el aprendizaje automático", en Multiple Classifier Systems. Berlín, Alemania: Springer, 2000.

[25] L. McInnes, H. John y M. James, "Umap: Aproximación y proyección uniformes de múltiples para la reducción de dimensiones", 2018, arXiv: 1802.03426. [En línea]. Disponible: https://arxiv.org/abs/1802.03426 [26] T. Kyono, FJ Gilbert y M. van der Schaar, "MAMMO: Una solución de aprendizaje profundo para facilitar la colaboración entre radiólogos y máquinas en el diagnóstico de cáncer de mama", 2018 , arXiv:1811.02661. [En línea]. Disponible: https://arxiv.org/abs/1811.02661 [27] E.-K. Kim et al., "Aplicación de biomarcadores de imágenes basados en datos en mamografía para la detección del cáncer de mama: estudio preliminar", Sci. Rep., vol. 8 de febrero de 2018, art. No. 2762.

[28] W. Zhu, Q. Lou, YS Vang y X. Xie, "Redes profundas de instancias múltiples con asignación de etiquetas dispersas para la clasificación de mamografías completas", en Proc. MICCAI, 2017, págs. 603–611.

[29] T. Kooi y N. Karssemeijer, "Clasificación de diferencias simétricas y cambio temporal para la detección de masas malignas en mamografía utilizando redes neuronales profundas", Proc. SPIE, vol. 4, núm. 4, 2017, art. No. 044501.

[30] T. Kooi et al., "Aprendizaje profundo a gran escala para la detección asistida por computadora de lesiones mamográficas", Med. Image Anal., vol. 35, págs. 303–312, enero de 2017.

[31] L. Shen, "Entrenamiento de extremo a extremo para el diagnóstico de cáncer de mama de imagen completa utilizando un diseño totalmente convolucional", 2017, arXiv: 1711.05775. [En línea]. Disponible: https://arxiv.org/abs/1711.05775 [32] P. Teare, M. Fishman, O. Benzaquen, E. Toledano y E. Elnekave, "Detección de malignidad en mamografía utilizando redes neuronales convolucionales profundas duales y genéticamente descubrió una mejora de la entrada de color falso", J. Digit. Imag., vol. 30, núm. 4, págs. 499–505, 2017.

[33] Y. Gao, KJ Geras, AA Lewin y L. Moy, "Nuevas fronteras: una actualización sobre el diagnóstico asistido por computadora para imágenes mamarias en la era de la inteligencia artificial", Amer . J. Roentgenol., vol. 212, núm. 2, págs. 300–307, 2018.

[34] H. Harvey et al., "El papel del aprendizaje profundo en la detección de mamas", Informes actuales sobre el cáncer de mama, vol. 11, núm. 1, págs. 17 a 22, 2019.

[35] AS Becker, M. Marcon, S. Ghafoor, MC Wurnig, T. Frauenfelder y A. Boss, "Aprendizaje profundo en mamografía: precisión diagnóstica de un software de análisis de imágenes multipropósito en la detección del cáncer de mama" Investigative Radiol., vol. 52, núm. 7, págs. 434–440, 2017.

[36] A. Rodríguez-Ruiz et al., "Inteligencia artificial independiente para la detección de cáncer de mama en mamografía: comparación con 101 radiólogos" J.Nat. Cancer Inst., vol. 111, núm. 9, págs. 916–922, 2019.

[37] A. Rodríguez-Ruiz et al., "Detección de cáncer de mama con mamografía: efecto de un sistema de apoyo de inteligencia artificial", Radiología, vol. 290, núm. 2, págs. 305–314, 2018.

[38] IC Moreira, I. Amaral, I. Domingues, A. Cardoso, MJ Cardoso y JS Cardoso, "INbreast: Hacia una base de datos mamográfica digital de campo completo", Acad. Radiol., vol. 19, núm. 2, págs. 236–248, 2012.

[39] M. Heath et al., "Estado actual de la base de datos digital para la mamografía de detección", en Digital Mammography. Dordrecht, Países Bajos: Springer, 1998.