A DEEP NEURAL NETWORK FOR DESCRIBING BREAST ULTRASOUND IMAGES IN NATURAL LANGUAGE

Trabajo Fin de Máster

Autor:

Mikel Carrilero Mardones

Tutores:

Alberto Nogales Moyano

Jorge Pérez Martín

Francisco Javier Díez Vegas

Máster Universitario en Investigación en Inteligencia Artificial

Septiembre, 2022

Universidad Nacional de Educación a Distancia (UNED)
Escuela Técnica Superior de Ingeniería Informática

Índice general

Resumen

El cáncer de mama es el tipo de cáncer más común y la principal causa de mortalidad en la población femenina. Sin embargo, su detección temprana puede incrementar la tasa de supervivencia relativa a cinco años del 29 % al 99 %. La ecografía es una de las técnicas más utilizadas para el diagnóstico de cáncer de mama, pero es necesario un experto para interpretar sus resultados de forma correcta. Esto no es común en algunos países que no cuentan con un programa de cribado apropiado, suponiendo una bajada de la tasa al 20 %. Los diagnósticos asistidos por ordenador (CAD) tratan de ayudar a los médicos en este proceso, mejorando los resultados y ahorrando tiempo. Los expertos en cáncer de mama emplean la clasificación BI-RADS para describir tumores, estimar su malignidad y establecer el tratamiento a seguir. Mientras la mayoría de sistemas CAD se limitan a clasificar imágenes según su malignidad, presentamos un modelo basado en dos sistemas para la detección y descripción en lenguaje BI-RADS de tumores en tiempo real. El primer sistema es un algoritmo de detección basado en YOLO que obtiene una precisión de 0.965, una exhaustividad de 0.95 y un área bajo la curva de precisión-exhaustividad de 0.97. El segundo es un sistema de descripción que recibe el tumor detectado y devuelve, en lenguaje natural, su descripción en BI-RADS y una estimación de su malignidad. Para este sistema hemos realizado tres experimentos en colaboración con una radióloga experta en mama y hemos obtenido unos valores de concordancia con sus diagnósticos que se encuentran entre los valores de intercorrelación e intracorrelación entre expertos que hemos encontrado en la literatura. Además, observamos que entrenar los modelos con los descriptores BI-RADS mejora la clasificación según malignidad y los acerca al razonamiento experto.

Palabras clave

cáncer de mama, BI-RADS, aprendizaje profundo, descripción de imágenes médicas, ecografía, diagnósticos asistidos por ordenador, mecanismo de atención, inteligencia artificial explicable.

Abstract

Breast Cancer is the most common cancer and the first cause of mortality among female population. However, its early detection can increase the 5-year survival rate from 29% to 99%. Ultrasound is one of the most used techniques for breast cancer diagnosis, but an expert in the field is necessary to interpret the test correctly. This is not common in some countries that can not afford a proper screening program, resulting in a drop in the 5-year survival rate to 20%. Computer Aided Diagnosis (CAD) systems aim to help physicians during this process, improving results and saving time. Breast cancer experts use Breast Imaging-Reporting and Data System (BI-RADS) classification to describe tumors, estimate their malignancy and establish the treatment to follow. While most CAD systems are limmited to classifying ultrasound images as benign or malignant, giving an explanation via a Region of Interest or an attention mechanism, we have developed a two-system-based model for real-time tumor detection and description using BI-RADS language. The first system is a YOLO-based detection algorithm, which obtains a precision of 0.965, a recall of 0.95, and an area under the precision-recall curve of 0.97. The second is a description system, which uses detected tumor and outputs, in natural language, its description in BI-RADS, and an estimation of the malignancy. For this system, we have carried out three different experiments in collaboration with an expert radiologist in breast cancer and obtained agreement values with her diagnoses that lay between expert intercorrelation and intracorrelation. We also show how training the models with BI-RADS descriptors improves malignancy classification and brings the model closer to expert reasoning.

Keywords

breast cancer, BI-RADS, deep learning, medical image captioning, ultrasound image, computer aided diagnosis, attention mechanism, explainable artificial intelligence.

1 Introducción

El cáncer de mama fue el tipo de cáncer más común en 2018, tras el de pulmón, y fue el mayoritario en el año 2020, con aproximadamente 2.09 y 2.30 millones de casos en el mundo, respectivamente [1]. La tasa de supervivencia relativa a 5 años tras el diagnóstico es de aproximadamente el 90 %, y si este se detecta en la etapa temprana sube al 99 %. Lamentablemente, tan solo el 65 % de los cánceres de mama se detectan en esta etapa; el porcentaje baja al 47 % en las mujeres entre 15 y 39 años. Esta tasa se mide comparando la proporción de personas que han sobrevivido tras 5 años con esta enfermedad sobre la tasa de supervivencia de la población global. Así, la detección temprana de tumores malignos de mama es de vital importancia. Si el cáncer se expande de forma regional (a zonas cercanas a la mama), la tasa de supervivencia es del 86 % y, si se expande más allá, esta baja al 29 %. Estos números empeoran en algunos países como Mali y Gambia, donde la tase de supervivencia relativa a 5 años es del 20 % [2]. Esto se debe, en parte, a la falta de equipamiento y expertos necesarios para realizar un cribado correcto.

Existen diferentes técnicas para el diagnóstico de cáncer de mama, como la ecografía, la mamografía y la resonancia magnética. Aunque la mamografía es el estándar, por ser la más coste-efectiva, tiene sus desventajas. La paciente se ve expuesta a radiación que puede incrementar las probabilidades de contraer cáncer [3]. Además, la exhaustividad (también conocida como sensibilidad) es más baja en mamas densas, más comunes en jóvenes [4]. Por último, es necesario comprimir la mama, lo que puede causar dolor [5].

Por otro lado, la ecografía de mama es una herramienta indolora, no invasiva, barata y capaz de visualizar nódulos que pueden pasar inadvertidos en otras técnicas o ayudar a clarificar ciertas características de un tumor. Normalmente, la ecografía es utilizada junto a la mamografía o a la resonancia magnética como un método complementario [6]. Un problema de la ecografía es el bajo valor de la relación señal-ruido, que dificulta la extracción de información a los médicos y requiere de expertos en la técnica. Por ello, el diagnóstico asistidos por ordenador (CAD, por sus siglas en inglés) ha obtenido una gran importancia. Estos modelos tratan de ayudar al médico en la detección (CADe) y en la clasificación[1] (CADx), reduciendo su trabajo y mejorando el resultado final [7].

En todos los ámbitos médicos la estandarización de la sistematología y terminología para describir los hallazgos realizados es indispensable. El caso del cáncer de mama

[1] Durante este trabajo se hablará tan solo de clasificación para expresar la clasificación según la benignidad o malignidad del nódulo.

no es diferente y cuenta con un método universal propuesto por el *American College of Radiology*, el sistema BI-RADS (*Breast Imaging Reporting and Data System*) [8]. Este incluye una serie de descriptores como la forma, el margen, la orientación y el patrón ecogénico del nódulo, para finalmente dar un valor a la peligrosidad del tumor en una escala no lineal de 9 niveles. De esta manera, el médico sabrá cual es el procedimiento a seguir según el nivel de peligrosidad dado por el sistema. Un valor BI-RADS 0 indica que el examen no es concluyente y que debe repetirse. BI-RADS 1 indica que no se han encontrado hallazgos y que se procede a un seguimiento habitual. En BI-RADS 2 se han encontrado hallazgos benignos e igualmente se continúa con el seguimiento habitual. Con BI-RADS 3 los hallazgos son probablemente benignos, con una probabilidad menor al 2% de malignidad, y se realiza un control al de 6 meses. A partir del BI-RADS 4A se realiza una biopsia del nódulo para su clasificación. BI-RADS 4A indica que es un nódulo de baja sospecha y su probabilidad de malignidad es menor al 10%; BI-RADS 4B, de sospecha intermedia, entre 10% y 50%; BI-RADS 4C, de alta sospecha, entre 50% y 90%, y BI-RADS 5, sugestivo de carcinoma, mayor al 90%. Por último, tenemos el BI-RADS 6, que indica que el nódulo ya ha sido biopsiado y se ha clasificado como cancerígeno. En la Figura 1 encontramos la descripción de un quiste simple empleando el sistema BI-RADS. Este tipo de tumor benigno puede encontrarse en el 90% de las mujeres y es clasificado como BI-RADS 2.

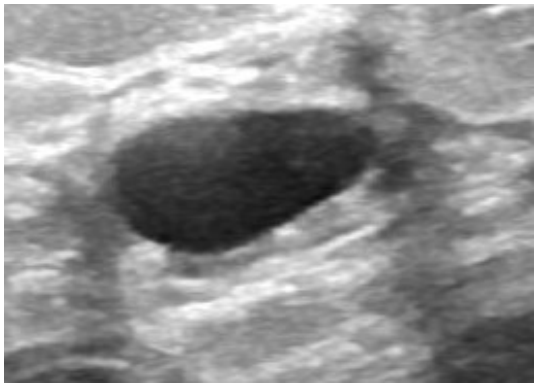

Figura 1 Descripción obtenida de la experta: *Nódulo ovalado, circunscrito, con orientación paralela y anecoico, con refuerzo posterior, sin halo ecogénico, sugestivo de quiste simple. BI-RADS 2.*

Sin embargo, esta clasificación del nódulo en características no es única y existe disparidad en la opinión de los expertos. Es decir, dos expertos pueden dar una descripción diferente en el sistema BI-RADS de un nódulo. Por ejemplo, puede haber disparidad en la forma del nódulo al elegir entre ovalado y redondeado. Esto ocurre también en el resto de características BI-RADS y se ha visto en mamografía [9], en ecografía [9, 10, 11], en la combinación de ambas [12] y en resonancia magnética [13]. En todos estos estu-

dios se emplea la métrica kappa de Cohen, que veremos en detalle en la sección 3.5. La mayor correlación entre expertos se da cuando estos tienen al alcance las imágenes obtenidas por ecografía y mamografía a la vez [12]. Al tratarse este trabajo sobre ecografía de mama, nos centramos en especial en los primeros estudios publicados [9, 10, 11]; en los dos últimos no solo observan la disparidad entre expertos (intercorrelación), si no la existente en un mismo doctor al analizar la misma imagen pasado un tiempo (intracorrelación). En la Tabla 1 pueden verse estos resultados. La media obtenida es una media ponderada teniendo en cuenta el número de nódulos analizados, esta será utilizada para comparar los resultados con los de nuestra red. En la intracorrelación, en uno de los trabajos se estudian los resultados de cuatro doctores, por lo que se ha hecho una media de ellos [11]. Puede apreciarse como en la mayoría de los casos el nivel de intercorrelación en expertos es moderada, con algún caso de correlación baja, el margen y la ecogenicidad. Asimismo, el nivel de intracorrelación en expertos es, en general, buena.

Tabla 1
Nivel de concordancia entre expertos [9, 10, 11]. Entre paréntesis el número de nódulos analizados.

	Forma	**Margen**	**Orientación**	**Ecogenicidad**	**Posterior**	**Frontera**	**Categoría**
Intercorrelación							
[9] (91)	0.66	0.4	0.61	0.29	0.4	0.69	0.28
[10] (314)	0.42	0.32	0.61	0.36	0.53	0.55	0.49
[11] (150)	0.49	0.33	0.56	0.37	0.49	0.59	0.53
Media	**0.48**	**0.34**	**0.6**	**0.35**	**0.5**	**0.58**	**0.47**
Intracorrelación							
[10]	0.73	0.64	0.68	0.65	0.64	0.68	0.74
[11]	0.66	0.57	0.8	0.74	0.72	0.74	0.75
Media	**0.68**	**0.59**	**0.76**	**0.72**	**0.69**	**0.72**	**0.75**

En este entorno se establece el trabajo de fin de máster que se elabora a continuación. Durante la investigación he construido un sistema CAD capaz de ayudar al médico tanto en la detección como en el diagnóstico de masas en ecografía de mama. Además, este modelo obtiene las características del sistema BI-RADS de cada nódulo detectado y devuelve una descripción en lenguaje natural, similar al que los radiólogos emiten en sus informes. No se ha encontrado otra arquitectura en la bibliografía que ayude al médico tanto en la detección de tumores, como en la redacción del informe médico en cáncer de mama. Esta red puede dividirse en dos sistemas esenciales, el detector y el descriptor. Para el primer sistema hemos empleado el conocido algoritmo YOLO (You Only Look Once) [14] que, mediante el uso de redes neuronales convolucionales (CNN), devuelve una región de interés (ROI, por sus siglas en inglés) de cada tumor detectado. Para el segundo sistema hemos generado dos modelos, es aquí donde entra

la parte más innovadora del trabajo y donde se centra el mayor esfuerzo. Cada uno de estos modelos devuelve una descripción del tumor extraído por YOLO en lenguaje natural y una estimación de su benignidad. Finalmente, se le ha añadido al modelo un algoritmo de explicación como Shapley Additive exPlanation (SHAP) [15] que indica en que características BI-RADS se ha basado el modelo para dar la salida de la clasificación final BI-RADS.

Este trabajo se inscribe dentro del proyecto "Cribado coste-efectivo del cáncer de mama mediante mamografía, ecografía y termografía", financiado por el Ministerio de Ciencia e Innovación (ref. PID2019-110686RB-I00), en el cual soy investigador. El estudio ha sido aceptado como póster en el congreso anual de EUSOBI (European Society of Breast Imaging) 2022, que se celebrará del 29 de septiembre al 1 de octubre en Malmö, Suecia.

1.1 Implicaciones Éticas

En esta sección se analizan las posibles consecuencias éticas y sociales del trabajo de investigación llevado a cabo. Para ello, se responden las preguntas pertinentes de las directrices para una IA fiable de la Comisión Europea [16]. Para hablar de las implicaciones éticas, tendremos en cuenta como se quiere implantar nuestro modelo en la práctica, siendo un método de ayuda en detección, descripción y clasificación de tumores en vídeo.

En primer lugar, hemos diseñado el modelo para que sea utilizado por un doctor y no trata de suplantarlo, sino de ser una ayuda para este. Es el propio doctor el que guía la ecografía, mientras que el modelo ayuda en la detección de nódulos, sin embargo, quedaría en la decisión del experto si generar una descripción de estos nódulos o no. Una vez obtenida la descripción, el doctor es capaz de cambiar las características del sistema BI-RADS para generar una nueva salida. Es decir, se trata de un modelo de soporte donde el doctor tiene la última palabra.

En segundo lugar, la interpretabilidad de los datos es algo indispensable en medicina y se impone en Europa por ley a través del Reglamento General de Protección de Datos (GDPR, por sus siglas en inglés) [17]. En concreto, el GDPR en el artículo 22 expresa que un individuo tiene derecho a conocer información significativa sobre la lógica que hay detrás de una decisión automatizada. Así, nuestro modelo logra una gran interpretabilidad sobre el diagnóstico en ecografía, pues indica la ROI donde se ha localizado el tumor y genera una descripción de este en el sistema estándar BI-RADS junto a la clasificación final. Además, se ha añadido el algoritmo SHAP que indica de forma vi-

sual y fácil de entender cuáles han sido los descriptores más importantes a la hora de decidir el BI-RADS final del tumor.

Por último, en cuanto a la privacidad de datos, para el entrenamiento se han utilizado bases de datos públicas donde no existe ninguna forma de identificación del paciente. En caso de seguir entrenando el modelo con datos de ecografías obtenidas en un hospital, estas igualmente serían anónimas. Además, durante el uso del sistema, este no obtiene ni emplea ningún tipo de dato personal correspondiente al sujeto.

1.2 Objetivos

El objetivo principal de este trabajo es la generación automática de informes médicos a través de imágenes de ecografías de mama mediante la detección, descripción y clasificación de tumores. Nuestro modelo puede ser dividido en dos sistemas principales: en primer lugar, el sistema que detecta y obtiene la ROI de tumores y, en segundo lugar, el sistema que obtiene los descriptores BI-RADS, la clasificación y el resultado final en lenguaje natural.

Los objetivos específicos son:

- Investigar el estado del arte para iniciar el estudio con los mejores métodos para el proyecto (capítulo 2). En este caso, tenemos que encontrar los mejores métodos de detección, descripción de imágenes y clasificación.
- Generar una metodología idónea para el trabajo en cuestión (sección 1.3).
- Generar descripciones de las ecografías de las bases de datos públicas mediante la ayuda de una experta. Esta fase es indispensable para un buen funcionamiento del modelo y, en muchas ocasiones, el objetivo más complejo de cumplir.
- Buscar una buena arquitectura de red para el problema en cuestión. Uno de los objetivos principales del modelo es generar descriptores que el médico sea capaz de entender, por lo que es en este apartado de la arquitectura en el que este trabajo es más innovador y donde se pone una mayor atención.
- Obtener y comparar los resultados de los modelos con los encontrados en el estado del arte. Al estar nuestra red dividida en detección y descripción, tendremos resultados diferentes para cada apartado.

1.3 Metodología

Durante la investigación se ha aplicado la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) normalmente utilizada en investigaciones de minería de datos, que es válida también para este trabajo [18]. Esta metodología consta de 7 etapas que se muestran en la Figura 2. La Tabla 2 indica cómo se ha aplicado a este trabajo de investigación.

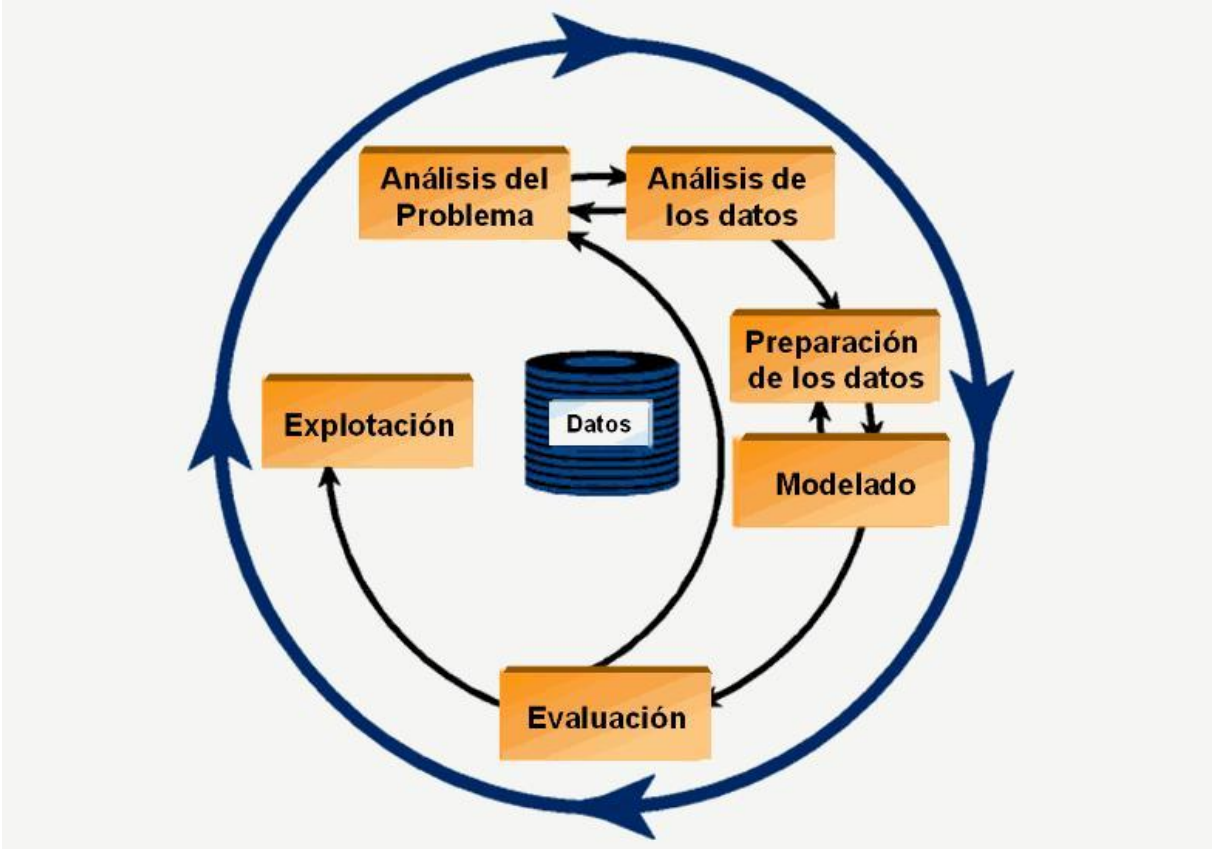

Figura 2 Siete etapas de la metodología Cross-Industry Standard Process for Data Mining (CRISP-DM) [18]. Una de las fases, iteraciones, está indicada por la flecha que se origina en "evaluación" y acaba en "análisis del problema". Imagen obtenida del estudio citado.

1.4 Estructura de Esta Memoria

El trabajo tiene una estructura especial, ya que contiene el artículo original que ha sido enviado a la revista *IEEE Transactions on Medical Imaging*.

Tabla 2
Procedimientos que se llevarán a cabo durante las etapas de la metodología CRISP-DM.

Etapa	**Procedimientos**
Análisis del Problema	• Estudio de modelos de detección. • Estudio de modelos de descripción de imágenes. • Estudio de modelos de clasificación de tumores de mama.
Análisis de los datos	Análisis de las bases de datos públicas que se emplean durante el trabajo y selección de imágenes para obtener descripción de la experta.
Preparación de los datos	Los sistemas de detección y descripción se entrenan por separado, por lo que los datos deben prepararse de diferente forma. • Se emplea la herramienta labelImg de Python para obtener las ROI de los tumores. • Se obtiene la descripción de los nódulos que la experta encuentra en las imágenes médicas y se convierten a valores numéricos (cambian según el modelo que empleemos). • Normalizar los valores y estandarizar la dimensión de las imágenes que servirán de entrada para el sistema de detección y descripción.
Modelado	Crear los sistemas adecuados para la detección y descripción considerando el número de imágenes médicas disponibles.
Evaluación	Evaluar los resultados utilizando las métricas pertinentes a cada sistema: • En el sistema de detección empleamos las métricas precisión, exhaustividad y AUPRC. • En el sistema de descripción empleamos exactitud y kappa de Cohen para los descriptores y, además, precisión, exhaustividad y F1 para la clasificación.
Iteraciones	Analizar los resultados y volver a la primera fase si estos no son buenos.
Explotación	Emplear nuestro sistema en hospitales con ecografías en tiempo real.

Además del artículo (capítulo 5), contamos con unos primeros capítulos introductorios donde hablaremos del estado del arte (capítulo 2) y del marco teórico (capítulo 3), necesarios para entender el trabajo que se ha realizado. Dentro del marco teórico encontramos las métricas empleadas en el artículo para la evaluación del modelo. También, hemos introducido un capítulo para explicar las iteraciones realizadas dentro de la metodología CRISP-DM (capítulo 4). Finalmente, en el capítulo 6 ofrecemos unas conclusiones y proponemos líneas futuras de investigación.

2 Problemática y Revisión de la Bibliografía

En medicina y, en concreto, en problemas relacionados con hallazgos mamográficos, los sistemas más habituales son los de segmentación, detección y clasificación de imágenes médicas, tanto en ecografía como mamografía. Comenzamos este capítulo estudiando la problemática más habitual de estos sistemas.

- Algoritmos de caja negra: Uno de los problemas más habituales en los modelos de visión artificial es que suelen ser algoritmos de caja negra, es decir, no dan una explicación de los resultados que obtienen. Como hemos visto en la sección 1.1, la interpretabilidad de los resultados es algo indispensable hoy en día, ya que el doctor debe ser capaz de entender la decisión del modelo y actuar al respecto. Este problema es muy habitual en modelos de clasificación, puesto que no es de gran ayuda que una red clasifique una imagen médica como benigna o maligna sin aclarar dónde ha visto este resultado.
- Falta de sistemas de descripción: Como hemos mencionado los sistemas más habituales son los de segmentación, detección y clasificación de imágenes médicas, mientras que los modelos de descripción en texto son minoritarios. Sin embargo, los expertos emplean los descriptores BI-RADS para dar un resultado de la posible malignidad de un tumor y actuar al respecto. Un modelo que siga estos mismos procesos ayudaría al médico a entender mejor el razonamiento tras los resultados propuestos y le ahorraría tiempo en la redacción de los informes.

Con los avances tecnológicos y el crecimiento de las bases de datos, la Inteligencia Artificial (IA) ha obtenido un gran impulso en las últimas décadas demostrando ser de gran utilidad en una gran variedad de problemas, como en reconocimiento de voz, minería de datos, Procesamiento del Lenguaje Natural (PLN) o visión artificial. Dentro de la IA, las redes neuronales, con sus diferentes tipos de arquitectura, son las que mejores resultados están obteniendo. La red neuronal más sencilla es la formada por una única neurona artificial, que es también conocida como perceptrón. El perceptrón toma las entradas, realiza una suma ponderada de estas y la introduce en una función no lineal conocida como función de activación. Los pesos de esta suma ponderada son aprendidos por la neurona a través de un proceso iterativo como el descenso por gradiente. Cada una de estas iteraciones a través de toda la base de datos es conocida como época. Las neuronas pueden agruparse formando capas y estructuras más complejas y, en caso de tener al menos una capa intermedia entre la de entrada y salida, estas redes son conocidas como profundas. El aprendizaje profundo ha mejorado de forma notable los resultados obtenidos por las tecnologías más modernas, gracias a su habilidad para

detectar características abstractas de los datos [19]. En el área de la visión artificial, los modelos que mejores resultados están obteniendo actualmente son las CNN [20]. Estas redes ya eran empleadas en 1995 para, por ejemplo, detectar cáncer de pulmón [21]. Sin embargo, fue tras la llegada de la red profunda Alexnet [22], que ganó el conocido concurso Imagenet en 2012, cuando estos modelos se convirtieron en estándar en técnicas de imagen. Así, los siguientes años nacieron otras CNN de uso habitual en medicina como VGG [23], Inception [24] o ResNet [25].

El gran problema de las redes neuronales profundas es que son modelos de caja negra. Por este motivo, los últimos años se ha hecho un gran esfuerzo por integrar técnicas que ayuden a interpretar la salida final de la red. La gran mayoría de estas técnicas suelen optar por la explicación visual y el número de artículos publicados crece cada año, mientras que la justificación en texto no es tan habitual (Figura 3) [26]. La explicación visual es una gran ventaja frente a los algoritmos de caja negra, dado que el experto cuenta con un método para saber donde ha puesto atención el modelo y así poder corroborar los resultados. De esta manera, estos algoritmos son capaces de señalar tumores posiblemente malignos de los cuales puede hacerse una biopsia. Una de las técnicas más recientes en explicación visual son los algoritmos de atención, que nacieron en el área del PLN como una ayuda a los sistemas *encoder-decoder* formados por redes *Long Short-Term Memory* (LSTM) [27] para no perder información al codificar frases de gran tamaño [28]. Sin embargo, su uso ha sido muy extenso y ha obtenido grandes resultados en el reconocimiento visual [29, 30]. Esto es debido a que estos algoritmos aprenden junto a la red y es sencillo visualizar sus resultados que marcan de qué parte de la imagen ha obtenido el modelo la información. En medicina también se están obteniendo grandes avances gracias a los modelos de atención; por ejemplo, se ha logrado mejorar los resultados de segmentación sobre las redes más empleadas en tres bases de datos correspondientes a tumores cerebrales, órganos abdominales y estructuras cardiovasculares [31]. Asimismo, el algoritmo de atención también ha sido muy utilizado en los sistemas de clasificación de tumores, donde se han propuesto, además, dos algoritmos de atención más complejos, la atención de dos capas y la atención *gated* [32].

Otra forma de explicación visual es optar por un sistema de detección de tumores junto al de clasificación, de manera que es posible indicar al experto en que región ha basado el resultado el modelo. En el área de detección de objetos el algoritmo que mejores resultados está obteniendo es YOLO [14]. Además, YOLO es una red completamente convolucional, lo que hace que pueda recibir imágenes de diferentes tamaños y logre resultados rápidos en inferencia. Así, esta técnica puede emplearse también para reconocimiento de objetos en vídeo. Desde su creación en 2015 se han generado

cinco versiones distintas de este algoritmo (una sexta versión se ha publicado durante la creación de este trabajo). Nosotros emplearemos la quinta, YOLOv5.

YOLO ha sido empleado en diversas ocasiones en trabajos relacionados con el cáncer de mama, donde, por ejemplo, con la versión YOLOv3 se han detectado el 89.4% de los tumores con un *Area Under the Precision-Recall Curve* (AUPRC) del 94.2% y 84.6% para clasificarlos según malignos y benignos, respectivamente [33]. De la misma forma, esta técnica también ha sido empleada en ecografía con buenos resultados [34]. El último trabajo citado compara los resultados obtenidos con diferentes técnicas de detección. Con el modelo YOLOv3 obtienen una precisión de detección del 96.89% y 96.58%, y una exhaustividad del 68.81% y 67.23% en tumores benignos y malignos, respectivamente.

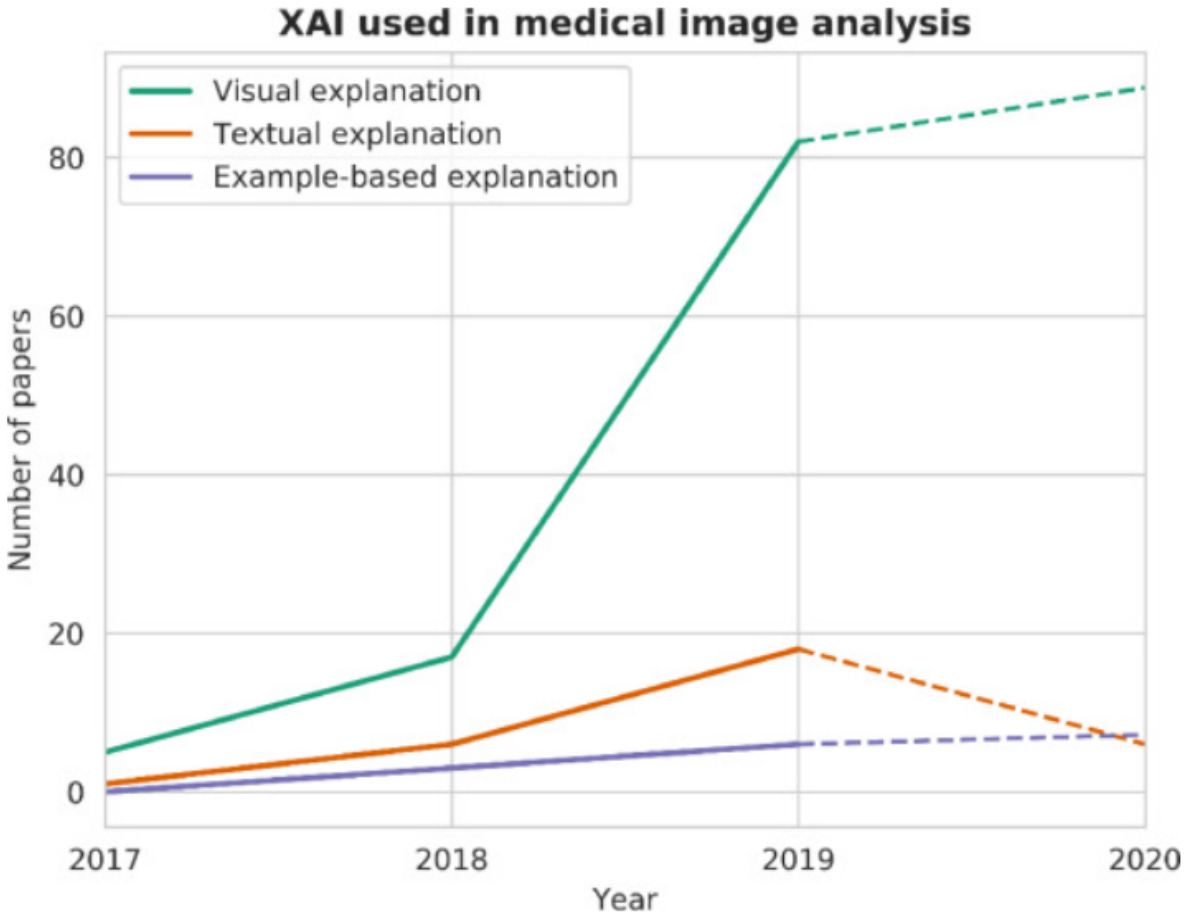

Figura 3 Imagen sacada de [26]. Número de artículos publicados en Inteligencia Artificial Explicable (XAI, por sus siglas en inglés.)

En cuanto a modelos de descripción de imagen los sistemas *autoencoder* (*encoder-decoder*) son los que mejores resultados están obteniendo. Los modelos *encoder-decoder*, como su

propio nombre indica, pueden ser divididos en dos sistemas. Primero, el modelo cuenta con un *encoder* que codifica la entrada y la convierte en propiedades abstractas de esta. Una vez la entrada ha sido codificada, esta pasa al sistema *decoder* que, a partir de estas propiedades, obtiene la salida de la red. Estas redes son muy empleadas, por ejemplo, en el PLN para problemas de traducción. El *encoder* convertirá la frase en un espacio de propiedades abstractas de esta, que luego será transformada por el *decoder* en el idioma deseado. Igualmente, este tipo de redes se ha aplicado en el ámbito de la visión artificial. Así, el objetivo del *encoder* en este caso es obtener las propiedades abstractas de la imagen.

En 2014 se creó un modelo que se componía de una CNN para extraer propiedades de las imágenes y una RNN para extraer las de las palabras [35]. En cada época se introduce en la RNN la palabra anterior y su salida se une con la de la CNN para usarla como entrada a una red densa con salida *softmax*, esta devuelve la siguiente palabra. En medicina este tipo de redes han sido empleadas para obtener descripciones de imágenes, por ejemplo, para describir imágenes de ecografía fetal obtenidas de vídeo [36]. El mayor problema de las redes de este tipo es el número de parámetros que contienen, ya que juntamos las propiedades de la CNN y la RNN de forma separada como entrada a una red densa. Además, esta red aunque da descripción en texto, no dice de qué parte de la imagen está obteniendo esta descripción.

En 2016 se creó un *autoencoder* que solucionaba los problemas anteriores [37]. Este emplea una LSTM en vez de una RNN y las propiedades obtenidas de la CNN (VGG16) son introducidas como entrada a la LSTM. Para ello, se emplea un algoritmo de atención sobre la imagen que toma como entrada también el estado oculto de la iteración anterior de la LSTM. Así, el sistema de atención devuelve una salida conocida como contexto, que contiene las propiedades más importantes de la imagen, y el anterior estado de la LSTM. Por último, la salida de la LSTM que toma como entrada este contexto devuelve directamente la siguiente palabra. Mediante esta arquitectura el número de parámetros disminuye, gracias al algoritmo de atención. En el ámbito médico existen algunos trabajos que han empleado este tipo de redes para generar descripciones a partir de las imágenes [38]. El estudio citado compara diferentes modelos empleando el conjunto de datos VQA [39], que está compuesto por diferentes tipos de imágenes médicas con descripción, y obtienen los mejores resultados con la arquitectura propuesta. Sin embargo, este no se centra en un área de la medicina concreta.

Existen otros trabajos donde emplean un método similar a la atención [38]; los resultados de las propiedades obtenidas por la CNN se introducen a otra CNN más pequeña que genera las ROI. De esta forma, son tan solo estas ROI las que se introducen en

la LSTM tras pasar por varias capas densas. La mayor diferencia es que la imagen se introduce a la LSTM tan solo para generar la primera palabra, mientras que los modelos de atención expuestos anteriormente tomaban el contexto como entrada en cada iteración. Igualmente, con este modelo obtienen mejores resultados en descripción de ecografía que los más utilizados (sin tener en cuenta los de atención) a partir de una base de datos de creación propia, con 3 órganos diferentes, 11 tipos de enfermedades y 4298 imágenes. Es cierto que estos modelos *autoencoder* pueden servir para detección y descripción de imagen al mismo tiempo, sin embargo, son notablemente más lentos que un modelo de detección como YOLO, ya que generan una descripción de cada imagen. Por lo tanto, estos sistemas no son viables para analizar vídeo.

Por último, existen varios trabajos que demuestran la eficacia de los descriptores BI-RADS para clasificar tumores de mama [40, 41]. En el primero de ellos se obtiene una precisión del 96.81 % en clasificación de 132 tumores tan solo a partir de estos descriptores. Sin embargo, estos son introducidos a mano, lo que conlleva un trabajo extra para el doctor. Otros trabajos que generan las características BI-RADS no tienen en cuenta si son correctas o no y las introducen directamente en una red para obtener el diagnóstico [42, 43]. La forma de obtener estos descriptores es a través de la diferencia en intensidad de píxeles para calcular las características relacionadas con la ecogenicidad y la segmentación para las relacionadas con la morfología. Igualmente, estos trabajos requieren un gran tiempo para un experto, ya que la segmentación debe hacerse manualmente. Tan solo hemos encontrado dos trabajos centrados en la descripción de ecografías de mama mediante el sistema BI-RADS que analizan los resultados obtenidos [44, 45]. El primero de ellos es un proyecto privado de la compañía Samsung, cuyo código no está disponible públicamente [44]. Sin embargo, en el trabajo citado se hace una comparación de los resultados del sistema con los de un experto, analizando 192 nódulos, empleando para ello la métrica kappa de Cohen. Los resultados pueden verse en la Tabla 3. Este modelo no dispone de sistema de detección y tan solo puede ser utilizado si el experto marca la ROI y los bordes del nódulo que quiere ser analizado. El segundo trabajo se publicó a la vez que se realizaba esta investigación y el método empleado para obtener los descriptores a través de la ROI es muy similar [45]. Esta investigación cuenta con la descripción de tres expertos sobre un total de 4458 imágenes de ecografía de mama que contienen un nódulo. Emplean un modelo de atención más complejo al aquí descrito [46]. Además, emplean diferentes redes para obtener cada descriptor. El objetivo principal de ese sistema es obtener las características BI-RADS que influyen de mayor forma en la clasificación de tumores para así poder ser introducidas en otra red. Por lo tanto, es un trabajo que no está pensado para la generación de informes médicos, ya que no recoge algunas de las características necesarias, pues solo busca obtener unos buenos descriptores para la clasificación final. Además, la ROI de

los tumores de mama se obtiene a través de los expertos y no por el propio sistema. Por último, no se demuestra si con los descriptores obtenidos se logran mejores resultados en clasificación que con los sistemas CADe actuales. En definitiva, es un trabajo puramente dedicado a lograr los descriptores BI-RADS más influyentes y comparar lo obtenido con la opinión de los expertos, con buenos resultados. Por ello, esta investigación se tendrá muy en cuenta en trabajos futuros para la creación de esta parte del sistema descriptor.

Tabla 3
Resultados de [44] en los descriptores BI-RADS.

	Forma	**Margen**	**Orientación**	**Ecogenicidad**	**Posterior**	**Frontera**	**Benignidad**
[44]	0.64	0.3	0.61	0.34	0.29	0.26	0.58

3

Marco Teórico

En esta sección vamos a revisar el marco teórico necesario para entender nuestro trabajo. Para ello, comenzaremos explicando las CNN y algunas de sus propiedades más importantes. En segundo lugar, introduciremos las redes LSTM, ya que uno de los modelos descriptores emplea esta arquitectura. En tercer lugar, explicaremos el funcionamiento del mecanismo de atención y sus ventajas. En cuarto lugar, hablaremos también de SHAP, ya que se ha empleado para obtener una mejor interpretabilidad de nuestro modelo. Por último, definiremos las métricas empleadas en nuestro estudio.

3.1 Redes Neuronales Convolucionales

Las redes neuronales convolucionales (CNN, por sus siglas en inglés) fueron creadas para simular la acción de las neuronas de la corteza visual primaria y han resultado muy eficaces para resolver problemas cuyas entradas tienen una topología en forma de rejilla, como series temporales o tareas de visión artificial. Nosotros nos centraremos en esta segunda área. Las imágenes están formadas por un espacio bidimensional que representa el alto y ancho de la imagen, y una tercera dimensión que indica sus propiedades. Al espacio bidimensional lo llamaremos representación y, a cada una de sus entradas, píxel o celda. La tercera dimensión puede expresar, por ejemplo, los colores de cada una de estas celdas. En caso de ser una imagen en blanco y negro, esta tercera dimensión será inexistente. En cada capa de la CNN la entrada y salida será una representación bidimensional, más una tercera dimensión con sus propiedades.

Empezamos por el caso más sencillo en el que la imagen esta formada tan solo por la representación. El primer elemento a definir en una CNN es el filtro o kernel, K. Este es una matriz bidimensional formado por valores reales, pero, cuyo tamaño es menor al de la imagen. Estos valores van a hacer la vez de pesos en las redes neuronales y son los parámetros a ajustar (puede verse un ejemplo en la matriz central de la Figura 4). Comenzando por la parte superior izquierda de la representación, realizamos una multiplicación elemento a elemento entre las entradas de la imagen y el filtro, hasta completar el tamaño del filtro, y sumamos este resultado, añadiendo un valor extra conocido como *bias*. A continuación, al valor obtenido se le aplica una función de activación para lograr la salida que se coloca en la posición $(1,1)$ de la que será la representación de entrada a la siguiente capa. El filtro se mueve en cada iteración para repetir el proceso, en primer lugar, en horizontal (tantas veces como sea posible hasta llegar al límite) y, en segundo lugar, en vertical, para volver al proceso horizontal. El número de píxeles que se desplaza el filtro en cada iteración se denota como *stride* y en este trabajo su valor es de uno, tanto de forma horizontal como vertical. En la Figura 4

puede verse como se aplica la convolución con un filtro 3×3 a una representación de tamaño 6×6, en la iteración número seis (cuatro veces de forma horizontal en la primera fila y dos más en la segunda).

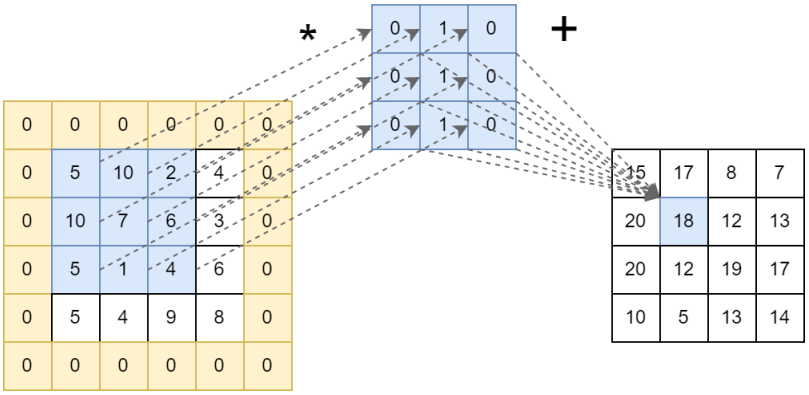

Figura 4 Convolución con filtro de tamaño 3×3 en la iteración 6.

Por cada filtro tenemos un total de parámetros igual a los píxeles que lo forman, más el *bias*, por lo que hay una gran diferencia entre las CNN y los perceptrones multicapa, puesto que estos últimos por capa tenían un total de pesos igual a la dimensión de la capa previa por la dimensión de la capa, $l_{prev} * l$. Esto provoca que las CNN tengan menos parámetros a ajustar, lo que agiliza mucho el proceso y evita el sobreajuste. Además, cada uno de estos filtros es capaz de detectar ciertos patrones a lo largo de toda la imagen, ya que sus parámetros no cambian al moverse por esta. Esto puede verse en la Figura 5, en la que se aplica un filtro capaz de detectar bordes horizontales. La primera capa de una CNN recibe como entrada la imagen tridimensional, donde la última dimensión indica los colores. para las imágenes en blanco y negro su tamaño será de uno, mientras que para las de color será de tres. Habiendo visto el caso simple en el que la entrada era tan solo la representación, pasamos a la entrada tridimensional. El filtro, en este caso, es también tridimensional y la multiplicación elemento a elemento se hace a través de las tres dimensiones. Nuevamente, sumamos este resultado y lo pasamos por la función de activación que dará la entrada $(1,1)$ de la representación de salida. Las capas CNN están formadas por más de un filtro, por lo que tendremos una representación diferente por cada filtro que apliquemos a la imagen. Una vez realizado el proceso con todos los filtros, se concatenan las representaciones para obtener la imagen tridimensional que servirá como entrada a la siguiente capa. Es decir, la tercera dimensión de la salida de la capa de la CNN es igual al número de filtros que tiene la capa. Por lo tanto, denotamos como filtros de entrada a la tercera dimensión que acompaña a las representaciones en la entrada de una capa (en la primera los colores) y filtros de salida a los filtros que se aplican en esta capa y que nos dará la tercera dimensión nuevamente para la siguiente capa. Este proceso puede ser expresado ma-

temáticamente a partir de la Ecuación 3.1, donde $Z_{j,k,i}$ es la salida de la capa antes de emplear la función de activación de la fila j, columna k y filtro de salida i; $V_{j,k,l}$ es la entrada a la capa de la línea j, columna k y filtro de entrada l, y $K_{m,n,i,l}$ es el filtro que se aplica a $\mathbf{V}$ con fila m, columna n, sobre el filtro de entrada i y con filtro de salida l.

$$Z_{j,k,i} = \sum_{m,n,l} V_{j+m-1,k+n-1,l} K_{m,n,i,l} \tag{3.1}$$

Figura 5 Convolución con filtro de tamaño 3×3 para detección de bordes horizontales.

Debido a este proceso, el tamaño de la representación disminuye de capa en capa obteniendo un valor de $\frac{n-f}{s}+1$, donde n es el tamaño de una dimensión de la representación anterior, f el del filtro y s el *stride*. Esto puede no ser un problema en CNN pequeñas, pero, debido a su gran eficiencia, es posible introducir un gran número de capas convolucionales, por lo que, en caso de no resolver este contratiempo, la dimensión de la representación podría reducirse drásticamente. Además, a las entradas que representan los vértices de la imagen tan solo se les aplica una vez los filtros de la capa CNN. Por estos motivos, se introduce el *padding* para aumentar la dimensión de la representación previo a introducirla como entrada en la capa posterior. En este caso, vamos a emplear el método llamado *zero padding*, que introduce nuevos píxeles con valor cero en la representación de entrada para lograr una salida de igual tamaño. El proceso de lograr una salida con mismo tamaño a la entrada es conocido como convolución *same*. En la Figura 4 puede verse como se ha aplicado un *zero-padding* de uno para volver a obtener una representación de 4×4.

En cada capa convolucional es habitual añadir otra función conocida como *pooling*. Esta se aplica sobre los píxeles cercanos de la salida obtenida por la capa convolucional y ayuda a convertir la representación en aproximadamente invariante a translaciones en la entrada, es decir, pequeñas traslaciones en la imagen no alteran el resultado de esta función. Además, estas funciones son utilizadas también para reducir la dimen-

sión. En este trabajo se va a emplear la función *max pooling*, que devuelve el máximo de una cuadrícula bidimensional; esta cuadrícula va a ser de tamaño 2×2. Debido a esto, la representación quedará reducida a la mitad por cada capa que empleemos. Un ejemplo de *max pooling* lo encontramos en la Figura 6, donde se aplica esta técnica a una entrada de tamaño 4×4, logrando una salida 2×2. Otra técnica de *pooling* es el *average-pooling*, que en vez de coger el máximo por cada cuadrícula, realiza una media de sus valores.

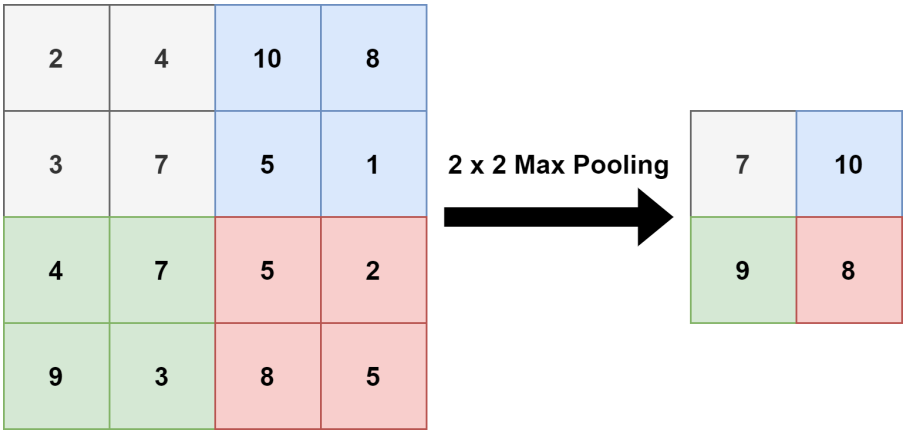

Figura 6 *max pooling* de 2×2 a una representación de 4×4.

Debido a la forma iterativa de aplicar el kernel, una ventaja de las redes totalmente convolucionales es que pueden tomar como entrada imágenes de diferente tamaño. Esto tendrá como resultado salidas también de tamaño variable, por lo que no es aplicable a todos los casos. Sin embargo, esto es de vital importancia para el algoritmo de detección empleado en nuestro trabajo, YOLO.

YOLO es un algoritmo capaz de detectar diferentes clases de objetos en una imagen. Para ello, el algoritmo devuelve la ROI por cada objeto detectado y lo clasifica según su clase. En el caso de las ecografías tan solo queremos que el algoritmo detecte una clase, nódulos. El algoritmo es una CNN cuya entrada es la imagen y su salida será una matriz tridimensional, donde las dos primeras dimensiones son una reducción del alto y ancho de la imagen. En la tercera dimensión tendremos la salida que marcará la probabilidad de que se encuentre un objeto en cada celda de la representación bidimensional, a que clase pertenece el objeto, donde se encuentra el centro de este objeto, y el alto y ancho de la ROI. Estas cinco propiedades son las que generan la tercera dimensión de salida, sin embargo, no estamos considerando el caso de que dos objetos se encuentren en la misma celda. Para que la red sea capaz de detectar ambos objetos, es necesario que el tamaño de esta última dimensión sea el doble. Así, podemos repetir el tamaño de esta dimensión tantas veces sea necesario, cada una de estas repeticiones

es conocida como *anchor*.

En la salida de una capa convolucional, cada una de las celdas de la representación se ha generado con una zona de la imagen original, de proporción igual a uno entre las dimensiones de la representación. La última dimensión es el número de filtros que tiene la capa, que tras el aprendizaje de la red, marcará las propiedades que la red detecta en la imagen en cada una de estas celdas de la representación. Una de las propiedades de las CNN es que, a medida que disminuimos la dimensión de la imagen original en la representación, aumenta el número de filtros utilizados y la complejidad de las propiedades que esta es capaz de encontrar en las celdas. Es bien sabido que las primeras capas de una CNN muestran propiedades sencillas como líneas verticales y horizontales, mientras que cuanto más profundo se llegue en la red, las propiedades que esta aprenderá serán más complejas. Así, YOLO disminuye el tamaño de la imagen hasta obtener una salida de $n \times n \times p$ donde n es la dimensión de la representación y p el número de filtros de esta capa, que marca las propiedades de la imagen. Por último, se le aplica una última capa convolucional con un kernel de tamaño 1×1 y número de filtros igual a cinco veces la cantidad de *anchors*, o número de objetos que podremos encontrar en conjunto en una celda. Esta es la salida final que describíamos en el anterior párrafo y que indica donde se encuentra cada objeto. Como mencionábamos anteriormente, al ser YOLO una red totalmente convolucional, la entrada a la red no tiene por qué ser de una dimensión predeterminada. Esto quiere decir que en la salida $n \times n \times p$, n tomará distintos valores dependiendo del tamaño de entrada. Sin embargo, al ser un algoritmo de detección donde cada celda indica si se encuentra un objeto o no, valores variantes de n no representan ningún problema. Es más, esto quiere decir que una vez entrenada la red, esta red puede emplearse en diferentes tamaños de imágenes o vídeos sin necesidad de ser reajustados.

3.2 LSTM

Las *Long Short-Term Memory* (LSTM) son un tipo complejo de Redes Neuronales Recurrentes (RNN, por sus siglas en inglés). Las RNN son un tipo de redes neuronales que se caracterizan por considerar la salida de las capas en instantes de tiempo anterior. Un ejemplo de una RNN puede verse en la Figura 7, donde la capa intermedia toma como entrada también la salida de la propia capa en la instancia anterior. Así, en cada instancia tenemos dos entradas x_t y h_{t-1} y una salida h_t. Este tipo de red tiene la ventaja de que acepta entradas de diferente longitud, por lo que es muy utilizada en el PLN. El problema es que la memoria de lo ocurrido en las instancias anteriores se va perdiendo a medida que avanzamos. Por lo que la red en las instancias finales no tendrá información de las entradas en las instancias iniciales. Por ejemplo, si la red

está generando la siguiente frase: "La chica se levantó del asiento, ...", debe ser capaz de guardar la información sobre el género del sujeto, ya que si a continuación escribe "él", la frase perderá el sentido.

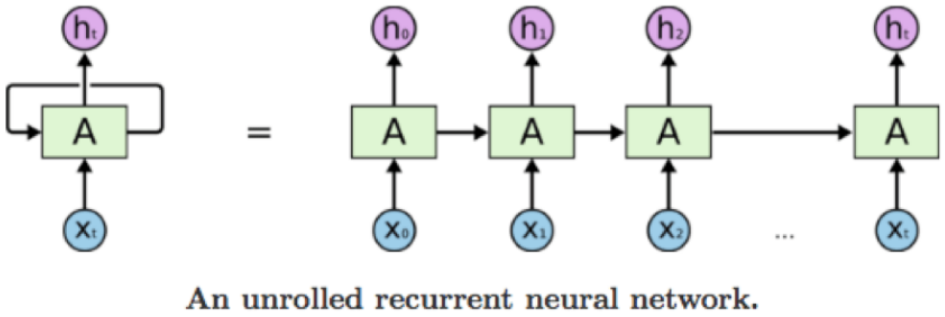

Figura 7 RNN con una sola capa intermedia recurrente. Imagen obtenida de https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Con el objetivo de solventar este problema se crearon las LSTM. Estas cuentan con una memoria que les da la capacidad de guardar información a largo plazo. En la Figura 8 puede verse un ejemplo de lo que es una capa LSTM. A diferencia de la RNN donde tan solo se empleaba la función de activación *tanh*, las LSTM tienen una estructura interna con una mayor complejidad. La red, en cada instancia, devuelve dos salidas, h_t y C_t. h_t es la salida final de la red y es conocida como estado oculto, mientras que C_t es la celda de estado o memoria. El estado oculto se emplea para dar la salida final de la red, mientras que la memoria guarda información sobre lo ocurrido en las diferentes instancias. En primer lugar, h_{t-1} y x_t son la entrada a una capa con función de activación sigmoidal que indicará que parte de la información de la memoria se olvida mediante una multiplicación por puntos; esta sección es conocida como compuerta *forget*. La función de activación sigmoidal obtiene salidas entre 0 y 1, por lo que permite el acceso o el cierre a la información que se tenía en la memoria C_{t-1}. En segundo lugar, tras elegir que información de la memoria anterior se elimina, se llega a la compuerta *update*, donde se actualiza C con la información de la entrada actual para obtener C_{t+1}. Para ello, se toma como entrada nuevamente h_{t-1} y x_t en una capa sigmoidal y en una capa *tanh* y se multiplica la salida de ambas. Nuevamente, la capa sigmoidal indica la importancia de los nuevos elementos para añadir en la memoria actual. Esta salida se suma al resultado de la memoria tras haber pasado por la compuerta *forget* para obtener C_{t+1}. Finalmente, h_{t+1} se calcula tomando como entrada h_{t-1} y x_t a una capa sigmoidal y multiplicando el resultado por la salida de C_{t+1} tras pasar por una función *tanh* (para obtener los valores entre -1 y 1).

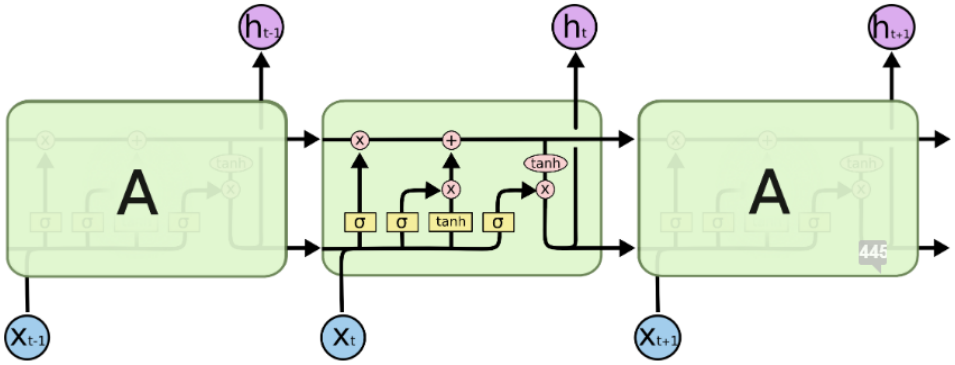

Figura 8 LSTM con una sola capa intermedia. Imagen obtenida de https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

3.3 Algoritmo de Atención

El algoritmo de atención nació en el área del PLN en los sistemas *seq2seq* o secuencia a secuencia [28]. Estos son *encoder-decoders* muy empleados en problemas como creación de resúmenes, chatbots o traducción de frases de un idioma a otro. Así, el *encoder* tomará la frase a traducir y la codificará. Una vez obtenida su codificación, esta pasa al sistema *decoder* que otorga la traducción en el otro idioma. Estos dos sistemas están formados por redes recurrentes como las LSTM. El problema es que en frases largas la codificación del sistema *encoder*, que trata del estado oculto y memoria de la última palabra de la frase a traducir, se va perdiendo a medida que avanzamos en las palabras del sistema *decoder*. Con el objetivo de no perder esta información se crearon los métodos de atención.

El algoritmo de atención emplea en cada iteración del sistema *decoder* todas las salidas de las iteraciones del *encoder*. Esto se hace mediante una media ponderada de ellas, donde los pesos indican la importancia que tiene cada salida en la iteración actual del *decoder*. Esta media ponderada conocida como contexto se introduce junto a la palabra de entrada en el *decoder*. Así, el sistema *decoder*, que se encarga de la traducción, tendrá información de todas las palabras codificadas por el sistema *encoder* y ponderadas por su importancia. Esta importancia de cada palabra del *encoder* se obtiene a través de un perceptrón que se entrena junto a la red. La entrada a este perceptrón es la salida de la palabra de la que queremos calcular la importancia en el *encoder*, más el estado oculto de la iteración anterior del *decoder*. El anterior estado oculto se añade para que el perceptrón obtenga información del sistema *decoder*. En caso de no emplear el estado oculto, la entrada al perceptrón sería la misma en cada iteración del *decoder* y, por lo tanto, también su salida. Esto indicaría que, para todas las palabras del sistema *decoder*, la importancia de las palabras del *encoder* es la misma. Una vez tenemos la im-

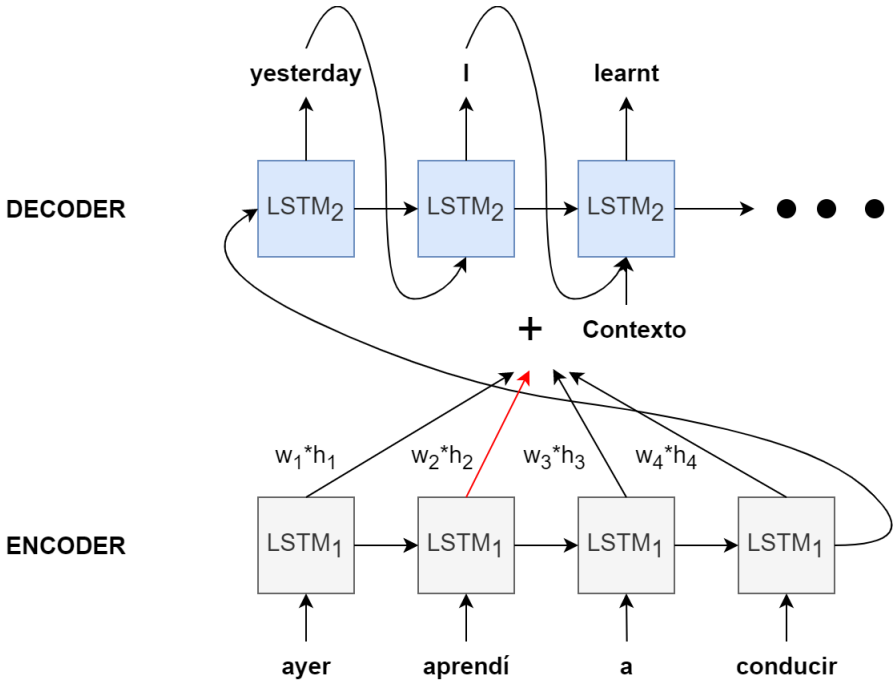

Figura 9 Algoritmo de atención en un sistema *encoder-decoder* de traducción mediante LSTM. El algoritmo muestra la atención en el momento de decodificar la palabra "learnt". h_i es el estado oculto en la iteración i del encoder. El peso w_i es el obtenido por el mecanismo de atención y muestra la importancia de la palabra i del encoder en esta iteracción del decoder. Cabe esperar que el mayor peso pertenezca a la salida de la palabra "aprendí", w_2.

portancia de todas las palabras del *encoder*, se emplea la función de activación *softmax* para obtener los pesos que se emplearán en la suma ponderada. En la Figura 9 puede verse un ejemplo de un sistema *encoder-decoder* con mecanismo de atención. En este caso, el sistema busca traducir la frase "ayer aprendí a conducir" y se encuentra en la tercera iteración del decoder tras obtener "yesterday I". Vemos que, gracias al mecanismo de atención, todas las palabras del *encoder* son consideradas, acompañadas de sus respectivos pesos, y la salida de "aprendí" es la que mayor importancia ha obtenido.

Aunque el algoritmo de atención nació en el PLN, pronto obtuvo protagonismo en el área de la visión artificial. En este caso, el *encoder* es una CNN que codifica la imagen obteniendo un espacio de propiedades de esta, de dimensión $n \times n \times p$. Recordamos que la última dimensión de este espacio contiene las propiedades de cada una de las celdas del espacio bidimensional $n \times n$. De este modo, cada una de las celdas hace de palabra en los sistemas *seq2seq*. Por lo tanto, en cada iteración del *decoder*, se introduce como entrada este espacio de dimensión $n^2 \times p$ junto al anterior estado oculto de la LSTM en el algoritmo de atención, y se obtiene una salida de la media de las importancias de estas celdas. Esta salida es el contexto, que en este caso tendrá dimensión p. Este es

el algoritmo de atención empleado en el primer sistema descriptor de nuestro trabajo, que hemos llamado *LSTM decoder*. Otra forma de algoritmo de atención en el área de la visión artificial, sin el uso de las LSTM, sería simplemente introducir el espacio $n^2 \times p$ en el algoritmo de atención, sin el estado oculto. De esta forma, obtendríamos una única salida de dimensión p, que puede emplearse como entrada en una red densa. Esto es similar a emplear el *average-pooling* sobre la última capa convolucional con una cuadrícula de tamaño $n \times n$, con la diferencia de que los pesos de las celdas no son idénticos, sino que están ordenados por su importancia. Este es el algoritmo de atención empleado en el segundo sistema descriptor de nuestro trabajo, que hemos llamado *Attention decoder*.

Hasta ahora hemos definido el algoritmo de atención como un perceptrón, pero es posible también emplear un perceptrón de dos capas o la atención *gated* [32]. La diferencia de la atención *gated* es que se emplean dos primeras capas, una con función de activación *tanh* y otra sigmoidal, realizando una multiplicación elemento a elemento de sus salidas.

3.4 SHAP

La explicabilidad de un modelo es indispensable hoy en día para los distintos problemas que aborda la IA. Una forma de explicabilidad es la interpretabilidad del modelo. La interpretabilidad indica conocer la causa y el efecto de una acción. Por ejemplo, imaginemos que quiero saber si la comida quemada es buena para el estomago. Es posible generar un modelo cuya entrada sea si la comida estaba quemada o no, y cuya salida si dañó el estomago. Este modelo podría tener una estructura tan compleja como deseemos, pero, si la salida cuando la comida estaba quemada es siempre que hizo daño y cuando estaba normal no, este modelo tendrá una gran interpretabilidad. Hemos generado una definición de relación entre la comida quemada y el daño al estomago. Este problema era muy sencillo, pero, a medida que crecen las variables de entrada y las salidas, la interpretabilidad se pierde. Debido a esto, existen algoritmos que se centran en obtener la interpretabilidad del modelo, como LIME [47] o, el que mejores resultados está obteniendo actualmente, *SHapley Additive exPlanation* (SHAP) [15].

SHAP, es un algoritmo basado en teoría de juegos que indica la importancia de cada variable de entrada a la hora de obtener la salida del modelo. Es empleado en modelos complejos para obtener una explicabilidad de su funcionamiento, algo indispensable en las redes neuronales profundas. Emplea los valores Shapley de teoría de juegos que indican la contribución que cada uno de los jugadores producen en el juego. En este caso, los jugadores son sustituidos por las variables de entrada y el resultado del juego

por la salida del modelo. SHAP es un modelo de interpretabilidad local, es decir, no estamos dando una interpretabilidad del modelo de forma única y global, sino sobre unos valores de entrada y salida concretos. Volvamos con el ejemplo anterior y añadamos otra variable de entrada sin relación, como el pijama que llevaba en el momento de comer. Para calcular los valores Shapley necesitamos una entrada concreta, como "quemado" y "pijama corto". La salida de la red será que hizo daño al estomago. Ahora, si nosotros no supiésemos si el pijama corto afecta en esto o no, no tendríamos una gran interpretabilidad del modelo, ya que no sabríamos la causa exacta del daño. Sin embargo, con lo valores Shapley obtendríamos que, para este caso, la contribución total para obtener "daño en el estomago" habría sido de la entrada "quemado".

Para explicar como se obtienen los valores Shapley vamos a partir de un ejemplo. Supongamos que queremos ver cuanto dinero nos otorga el banco según nuestro sueldo, la edad y si tenemos aval o no. En primer lugar, generamos todos los conjuntos posibles de combinación entre variables. Contando con el conjunto vacío tenemos un total de 2^3. Generamos un total de modelos igual al número de combinaciones obtenidas, donde la diferencia entre los modelos es la capa de entrada que se ajusta a las variables de cada conjunto. Así, entrenamos 8 modelos diferentes y analizamos la salida que obtienen para, por ejemplo, el caso "1200", "25" y "con aval". En cada modelo introducimos las variables con las que ha sido entrenado y obtenemos su salida. En la Figura 10 puede verse un diagrama explicativo de este resultado. Cada flecha representa la diferencia del precio al eliminar o añadir una variable, esta es conocida como la contribución marginal de esa variable. El resultado del conjunto vacío es el valor base del modelo, la media de los precios de salida, y, por ejemplo, añadirle la variable sueldo con valor 1200 disminuye el precio en 1000 euros. Esta es la contribución marginal del sueldo al conjunto vacío. Para calcular la contribución total del sueldo es necesario hacer una media ponderada de todas las contribuciones de esta variable (todas las flechas en dirección izquierda de la Figura 10). Para calcular los pesos tenemos las siguientes normas: la suma de estos pesos por fila en el diagrama debe ser la misma y los pesos que se encuentren en la misma fila tendrán también el mismo valor (esto puede verse nuevamente con claridad en la Figura 10). Por lo tanto, la fórmula para el cálculo de los pesos de la contribución marginal de una variable a un modelo con conjunto de variables f es $1/(|f| * \binom{F}{|f|})$, donde F es el número de variables total. Por ejemplo, en el primer caso donde tan solo hemos añadido la variable sueldo al vacío, f sería {"sueldo"} y tendríamos, $1/(1 * \binom{3}{1}) = 1/3$. De esta forma, la contribución final y valor Shapley de la variable sueldo para este ejemplo sería: $1/3 * (30{,}000 - 25{,}000) + 1/6 * (33{,}000 - 31{,}000) + 1/6 * (40{,}000 - 38{,}000) + 1/3 * (42{,}000 - 37{,}000) = -4000$

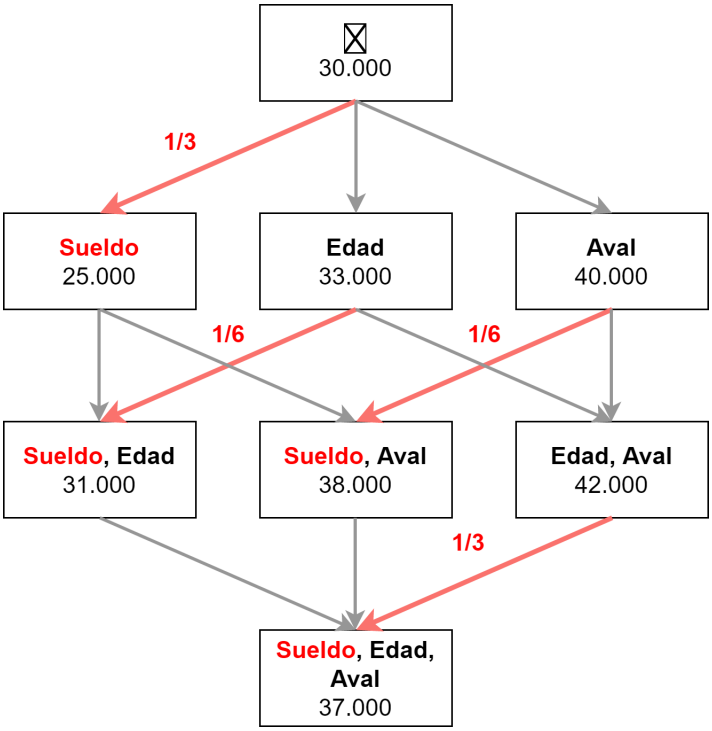

Figura 10 Diagrama de los 8 modelos entrenados con cada conjunto de variables y los resultados obtenidos. En rojo se marca la variable sobre la que se está calculando los valores Shapley, "Sueldo".

Sin embargo, para realizar este calculo es necesario entrenar 2^F modelos. Por este motivo, se emplean aproximaciones para obtener estas contribuciones. En este trabajo se ha empleado la aproximación KernelSHAP. En caso de querer calcular, por ejemplo, el resultado del conjunto "sueldo" y "aval" con los valores anteriores, se introducirán estos valores junto a una serie de valores aleatorios de la variable "edad" al modelo. Estos valores aleatorios se obtienen de ejemplos de la base de datos. Así, se introducirían "1200", "30", "con aval"; "1200", "27", "con aval"; "1200", "21", "con aval"; etc. Cuanto mayor sea el número de ejemplos, de mayor forma reduciremos la dependencia de esta variable. Una vez tenemos las muestras para todas las variables, estas se ponderan mediante la fórmula de la ecuación 3.2.

$$\pi_x(f) = \frac{F-1}{\binom{F}{|f|}|f|(F-|f|)} \tag{3.2}$$

donde x índica el ejemplo sobre el que estamos haciendo los cálculos, "1200", "25", "con aval". Así, en el trabajo citado [15] se demuestra que mediante una regresión lineal ponderada con estos pesos se obtienen los valores Shapley.

3.5 Métricas

En nuestro artículo se mencionan seis métricas que definimos a continuación:

- Exactitud: proporción de aciertos del modelo.
- Precisión (P), también llamada"valor predictivo positivo": proporción de verdaderos positivos entre los casos clasificados como positivos por el modelo,

$$P = \frac{vp}{vp + fp} \tag{3.3}$$

  donde vp son los verdaderos positivos y fp los falsos positivos obtenidos del sistema. Los verdaderos positivos son los casos que, habiendo sido clasificados como positivos, realmente eran positivos; mientras que, los falsos positivos son clasificados como positivos cuando eran negativos.
- Exhaustividad (R), también llamada "sensibilidad"; en inglés, "recall" o "sensitivity": proporción de verdaderos positivos obtenidos por el modelo sobre el total de positivos en la base de datos,

$$R = \frac{vp}{vp + fn} \tag{3.4}$$

  donde fn son los falsos negativos obtenidos del sistema. Los falsos negativos son los casos que, habiendo sido clasificados como negativos, eran positivos.
- F1: la media armónica de la exactitud y la precisión,

$$F1 = 2 * \frac{P * R}{P + R} \tag{3.5}$$

- *Area Under the Precision-Recall Curve* (AUPRC), también conocida como *Average Precision* (AP): esta métrica es más compleja que las demás y, al igual que la F1, se logra a partir de la exhaustividad y precisión. Según el umbral que escojamos para definir si un caso es positivo o negativo, la precisión y la exhaustividad cambian. Cuanto más estricto sea el sistema (mayor el umbral) y tan solo devuelva casos en los que está más seguro de que son positivos, mayor será la precisión y menor la exhaustividad, y viceversa. Por lo tanto, se puede generar una gráfica donde el eje x marque la exhaustividad, el eje y la precisión y se analicen sus valores según este cambio en el umbral. Así, calculando el área que deja esta curva sobre los ejes, se obtiene la métrica AUPRC (Figura 11).
- Kappa de Cohen (κ): Muy empleada en el ámbito médico para obtener un valor de coincidencia entre expertos. Esta métrica tiene en cuenta la probabilidad de que estos coincidan en el diagnóstico de forma casual. En la Tabla 4 podemos ver

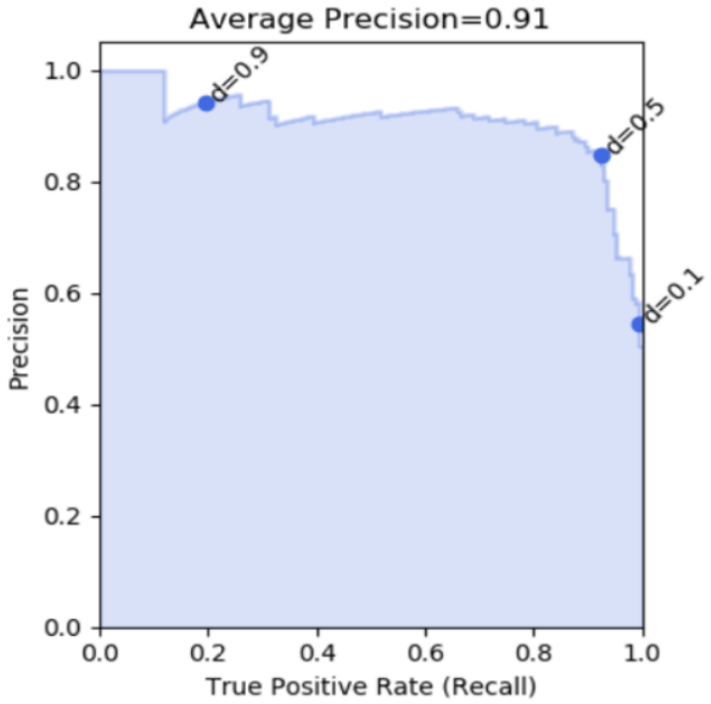

Figura 11 *d* marca el umbral. Imagen obtenida de https://glassboxmedicine.com/2020/07/14/the-complete-guide-to-auc-and-average-precision-simulations-and-visualizations/

el significado del resultado numérico de esta métrica,

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{3.6}$$

donde $Pr(a)$ es el porcentaje de concordancia entre ambos y $Pr(e)$ es la probabilidad de que esta se de de forma aleatoria.

Tabla 4
Nivel de concordancia entre dos expertos según el resultado kappa de Cohen.

Valor	Nivel de Concordancia
<0	Nada
0-0.2	Insignificante
0.2-0.4	Bajo
0.4-0.6	Moderado
0.6-0.8	Bueno
0.8-1	Muy bueno

4

Iteraciones en la Metodología y Primeros Pasos

En este trabajo presentamos dos modelos para el sistema descriptor. Ambos modelos son *encoder-decoders* y utilizan VGG16 como *encoder*. El sistema decoder del primero, *LSTM decoder*, está compuesto por una red LSTM con atención, mientras que el segundo, *Attention decoder*, es una red densa con atención. Hemos generado ambas redes empleando el paquete Keras de Python y creando nuestras propias capas. Ambas capas cuentan además con la opción de emplear la atención de doble capa y la *gated*. Estas atenciones han sido probadas sin una gran diferencia en los resultados.

Durante este trabajo hemos realizado varias iteraciones dentro de la metodología CRISP-DM. En un comienzo no contábamos con el sistema detector. Introducir YOLO a la red aumentó considerablemente los resultados por varios motivos. En primer lugar, el sistema descriptor tan solo tenía que extraer las características del nódulo a partir de la ROI, evitando así el ruido de la ecografía completa. En segundo lugar, nuestra red era capaz de detectar y describir diferentes nódulos en una misma imagen de forma separada, cuando antes describía la imagen completa. Además, en nuestra base de datos todas las ecografías que contuviesen más de un nódulo eran benignas. Si esto se le enseña a un modelo, puede ocurrir que este aprenda que la aparición de más de un nódulo sea sinónimo de benignidad (efecto Clever Hans [48]); sin embargo, es posible encontrar más de un tumor maligno o un tumor maligno cercano a unos benignos en una ecografía. Asimismo, la complejidad necesaria para que el sistema descriptor pudiera devolver características de más de un nódulo al mismo tiempo era demasiado alta para la base de datos que teníamos.

Probamos a aplicar dos filtros a las ecografías y generar una nueva imagen tridimensional con la mezcla de estas tres. El primer filtro era CLAHE [49], que sin entrar en detalles, aumenta el contraste en la imagen y suele ser empleado en mamografías [50]. El segundo era un filtro bilateral de creación propia que calcula el ruido de la imagen y ajusta sus parámetros según este, basándome en trabajos anteriores [51]. Sin embargo, los resultados que se obtenían eran muy similares a los de las imágenes en blanco y negro, por lo que se descartó esta alternativa. Aun así, estos filtros pueden ser interesantes para trabajos futuros, ya que cada uno muestra de forma más clara algunos de los descriptores. En la Figura 12 puede verse un ejemplo de estos dos filtros aplicados sobre una ROI.

Los dos modelos descriptores que presentamos en este trabajo dan la estimación de

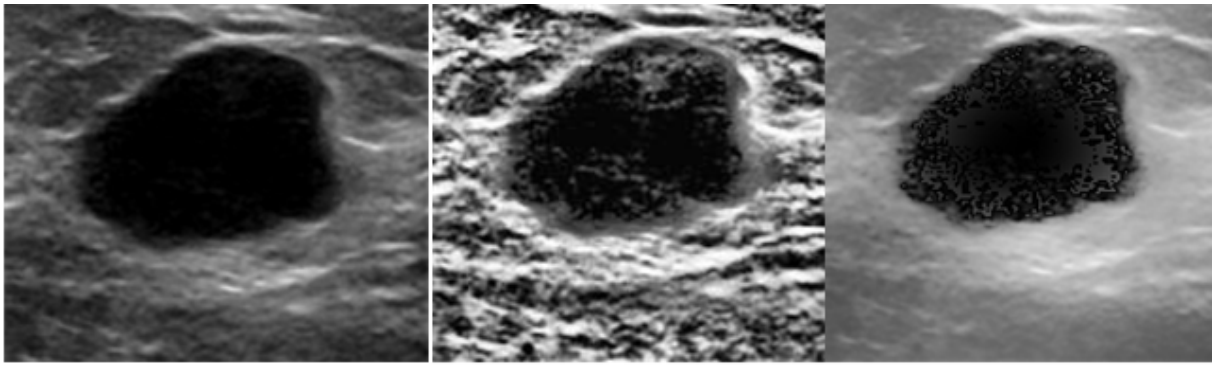

Figura 12 Los filtros CLAHE y bilateral aplicados a un nódulo. En el centro CLAHE y a la derecha bilateral.

malignidad de forma conjunta con las características BI-RADS. Probamos a separar este resultado del resto, ya que contábamos con más datos para entrenarlo (todas las imágenes de las distintas bases de datos, no solo de las que tenemos descripción). Sin embargo, los resultados obtenidos eran peores a los que teníamos entrenando la clasificación junto a las características. Esto puede deberse a que aprender los descriptores junto a la estimación de malignidad ayuda a la red en esta última tarea. Aun así, gracias a este procedimiento, nos dimos cuenta de un fallo grave en una de las bases de datos principales que se emplea a menudo en la detección y clasificación de ecografías de mama.

De las tres bases de datos públicas que se han analizado en este trabajo, una de ellas contiene un gran número de imágenes repetidas [52]. Estas repeticiones no son completamente idénticas, pero corresponden a la misma imagen sobre la que se han aplicado rotaciones o zoom. Para detectar las repeticiones utilizamos el algoritmo Space-Invariant Feature Transform (SIFT). SIFT detecta los llamados puntos de interés que permanecen invariantes a rotaciones y al escalado de las imágenes. Así, si el 50 % de estos puntos coincidían, estas imágenes se consideraron como idénticas. De esta manera, encontramos que de las 780 imágenes que componían la base de datos, 189 eran transformaciones. Además, 8 de estas transformaciones se encontraban en una clase distinta a la original. Se eliminaron, también, 26 imágenes de ecografía que pertenecían a la axila y 123 que contenían información como la ROI. Por lo que nos quedamos con un total de 442 imágenes. Cabe destacar que el algoritmo SIFT es restrictivo y no considera imágenes de la misma ecografía movidas un instante en el tiempo. Un ejemplo de un nódulo que aparecía tanto en benigno como en maligno lo tenemos en la Figura 13, además de un nódulo movido en el tiempo que SIFT no considera como repetición. Uno de estos primeros nódulos repetidos en ambas clases lo habíamos empleado como maligno hasta darnos cuenta del error. En *Google Scholar* hemos encontrado 205 trabajos que usan esta base de datos, pero ninguna de ellas menciona el

problema de las imágenes duplicadas. Las otras dos bases de datos también contenían repeticiones, aunque tan solo 2 [53] y 8 [54] fueron eliminadas, por ser transformaciones. Por estos motivos, hemos puesto mucha atención sobre las imágenes que hemos empleado para el trabajo, evitando copias e imágenes del mismo nódulo movidas en el tiempo.

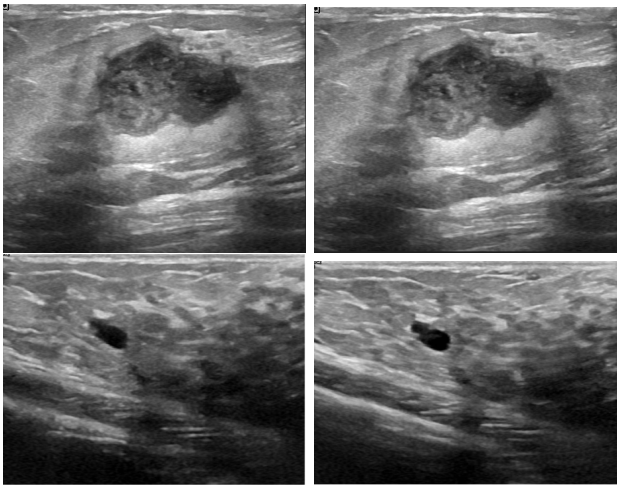

Figura 13 Las primeras dos imágenes fueron clasificadas como copias por SIFT, la primera pertenece a la clase "benigno" y la segunda a "maligno". Las últimas dos imágenes son el mismo nódulo movido en el tiempo (confirmado por la experta) y no fue clasificado como copia.

Hemos probado a entrenar VGG16 congelando distintas capas de la red en diferentes épocas. Finalmente, al examinar los resultados hemos comprobado que la última capa convolucional es la única que conviene entrenar desde cero. Asimismo, hemos probado a normalizar la salida del encoder VGG16 mediante la normalización Batch. Igualmente, debido a los resultados, esta tan solo se ha aplicado a la red *LSTM decoder*.

Por último, durante el proceso de creación de las arquitecturas, el número de nódulos con descripción BI-RADS ha ido aumentando a medida que se realizaban más reuniones con la experta. Es bien sabido que, cuanto mayor sea el conjunto de datos, más complejidad puede tener el modelo sin que este se sobreajuste durante el entrenamiento. Por ello, nuestros sistemas han ido cambiando a medida que se obtenían más descripciones.

5

Artículo

En este capítulo mostramos el artículo que ha sido enviado a la revista *IEEE Transactions on Medical Imaging* el día 28 de julio de 2022. En primer lugar, contamos con una sección de introducción en el que se habla también del estado del arte y los objetivos del trabajo. En segundo lugar, está la sección de métodos en la que describimos la arquitectura detallada de los modelos y en que consisten los diferentes experimentos. Además, se habla también de la preparación de los datos. En tercer lugar, encontramos la sección de resultados donde, mediante tablas, se enseña el rendimiento de las redes propuestas, comparándolas con otros modelos y los valores de concordancia entre los expertos. En cuarto lugar, tenemos la sección de discusión donde se ahonda en estos resultados, explicando los motivos y consecuencias. Por último, encontramos una breve conclusión de lo que ha sido la investigación y los trabajos futuros.

A Deep Neural Network for Describing Breast Ultrasound Images in Natural Language

Mikel Carrilero-Mardones, Manuela Parras-Jurado, Alberto Nogales, Jorge Pérez-Martín, and Francisco Javier Díez

***Abstract*— Breast Cancer is the first cause of cancer worldwide, and its early detection can increase the 5-year survival rate from 29% to 99%. Ultrasound is one of the most used techniques for breast cancer diagnosis, but an expert in the field is necessary to interpret the test correctly. This is not common in some countries that can not afford a proper screening program, resulting in a drop in the 5-year survival rate to 20%. Computer Aided Diagnosis (CAD) systems aim to help physicians during this process, improving results and saving time. Breast cancer experts use Breast Imaging-Reporting and Data System (BI-RADS) language to describe tumors, estimate their malignancy and establish a standard procedure. While most CAD systems focus on classifying ultrasound images as benign or malignant, giving an explanation via a Region of Interest or an attention mechanism, we have developed a two-system-based model for real-time tumor detection and description using BI-RADS language. The first system is a YOLO-based detection algorithm, which obtains a precision of 0.965, a recall of 0.95, and an area under the precision-recall curve of 0.97. The second is a description system, which uses detected tumor and outputs, in natural language, its description in BI-RADS, and an estimation of the malignancy. For this system, we have carried out three different experiments and obtained agreement values with our expert that lay between expert intercorrelation and intracorrelation. We also show how training the models with BI-RADS descriptors improves malignancy classification and brings the model closer to expert reasoning.**

***Index Terms*— BI-RADS, medical image captioning, ultrasound image, computer aided diagnosis, attention mechanism, explainable artificial intelligence.**

I. Introduction

Breast cancer was the second cause of cancer in 2018, and the first in 2020, with approximately 2.09 and 2.3 million cases, respectively [1]. A patient diagnosed with breast cancer has a 90% 5-year survival rate, and if it gets detected in an early stage, meaning it is localized, the survival rate rises to 99%[1]. If it has spread to regional lymph nodes, the survival rate is 86%; while if it has spread farther, it lowers to 29%. Only 65% of the cases are diagnosed in this first step, 47% for people between 15 and 39; therefore, breast cancer early detection is a primary concern. These numbers worsen in some countries, such as Mali and Gambia, where the overall 5-year survival rate is 20% [2]. This fact is partly due to not having the tests and experts necessary to implement screening programs successfully.

Different techniques exist for breast cancer diagnoses, such as ultrasound, mammography, or magnetic resonance. While mammography is the gold standard and most effective breast cancer screening test, it has some disadvantages. The patient is exposed to radiation that may increase the probability of developing cancer [3]. In addition, its recall is lower in dense breasts that are more common in young people [4]. Finally, breast compression is needed, which may cause pain to some women [5].

In contrast, ultrasound is a non-invasive, painless, and cheaper technique that is usually combined with mammography to verify results. It can detect nodules that may go unnoticed or clarify some specific tumor characteristics. The main problem with ultrasound is that images have a high noise-signal ratio and the presence of an expert is necessary to interpret the results. For these reasons, Computer Aided Diagnosis (CAD) systems have evolved in this topic to help radiologists during tumor detection (CADe) and classification (CADx) to save time and improve results [6].

With the technological advances and the growth of databases, Artificial Intelligence (AI) has proved to be a helpful tool to solve a wide variety of problems in different areas, such as speech recognition, data mining, Natural Language Processing (NLP), or computer vision. Deep learning, in particular, has dramatically improved state-of-the-art results due to its ability to obtain abstract characteristics from the data [7]. In computer vision, Convolutional Neural Networks (CNNs) have become the most effective models since AlexNet won the Imagenet competition by a significant margin in 2012 [8]. CNNs have fewer parameters per layer than feedforward neural networks and can learn space-invariant characteristics, allowing deeper architectures than

This paper was first submitted on august 2022. This work has been supported by grant PID2019-110686RB-I00 from the Spanish Government and grant 2022V/ITEMP/005 from the Universidad Nacional de Educación a Distancia (UNED).

Mikel Carrilero-Mardones, Jorge Pérez-Martín, and Francisco Javier Díez are with the Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain (e-mail:{mcarrilero, jperezmartin, fjdiez}@dia.uned.es).

Alberto Nogales is with the CEIEC Research Institute, Universidad Francisco de Vitoria, Madrid, Spain (email: alberto.nogales@ceiec.es).

Manuela Parras-Jurado is with the Department of Radiology, HM Hospitals, Madrid, Spain (email: mparras@hmhospitales.com).

[1] https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html

usual. Since 2012 all Imagenet winners have been CNN-based models, for example, Inception [9], Resnet [10], or VGG [11].

In medicine, there are numerous works in both CADe and CADx techniques, most of them with CNNs [12]. The main problem of these architectures is that they are black-box models, i.e., they are trained with extensive databases and learn patterns that may not follow human perception. These models are useless for physicians if they do not contain an algorithm that may give a perception of their inside. Physicians can not corroborate what the algorithm outputs and, hence, can not base their diagnosis on them. Moreover, the European General Data Protection Regulation (GDPR) establishes that explanations of machine learning techniques are mandatory [13]. For these reasons, eXplainable Artificial Intelligence (XAI) techniques have gained interest in recent years. Most of the current works in explainability for medicine use visual explanation, being text-based and example-based explanations less frequent [14]. Visual explanations focus on where the algorithm has put more attention and can return, for example, Regions of Interests (ROI), where the model has seen malignancy traces. This fact can help doctors to detect and conduct a biopsy on possible cancers. Nevertheless, they do not give a real inside of the final result.

The main motivation of this work is to generate a two-system-based model that not only detects and returns the ROI of an image but also gives a text-based explanation of it. Medical doctors diagnose tumors using Breast Imaging-Reporting and Data System (BI-RADS), a universal language for nodule description. This language contains descriptors, such as shape, margin, orientation, echogenicity, posterior enhancement, and echogenic halo. These descriptors are used to provide a class from the BI-RADS scale of the tumor malignancy (Table I). Few works are published in machine learning and BI-RADS descriptors. Some of them do not give a way to obtain them, but they use the descriptors as inputs to an algorithm to classify tumors according to their malignancy, with excellent results [15], [16]. Others, before inputting the descriptors to an algorithm, use pixel intensity techniques and handmade segmentation of the nodules to estimate them [17], [18]. Nonetheless, they do not measure the accuracy of descriptors and require extra work for the expert. We only found two researches focused on these measures [19], [20]. The first paper compares Samsung's S-detect system with an expert radiologist [19]. As a private corporation software, no information about the algorithm they used is available, but handmade detection and segmentation are again necessary. The second work uses deep neural networks with attention mechanisms to automatically obtain certain BI-RADS descriptors out of the ROI of a nodule [20]. In this work, they generate a dataset of 4,458 images described by three radiologists and use 10-fold cross-validation (CV) with different CNNs for not correlated descriptors. In this case, the only extra work is extracting the ROI from the image. Nevertheless, again, they focus on descriptors to input them into a network, given the results of previous works [15], [16]. Hence, they do not give echogenic halo characteristic and BI-RADS final classification.

TABLE I
BI-RADS CLASSIFICATION AND PROCEDURE TO FOLLOW ACCORDING TO THE MALIGNANCY OF THE TUMOR.

BI-RADS	Malignancy	Procedure to Follow
0	Incomplete	Additional evaluation
1	No findings	Normal procedure
2	Benign	Normal procedure
3	Probably benign ($< 2\%$)	Control in 6 months
4A	Low suspicion ($2\% - 10\%$)	Biopsy
4B	Medium suspicion ($10\% - 50\%$)	Biopsy
4C	High suspicion ($50\% - 90\%$)	Biopsy
5	Probably benign ($> 90\%$)	Biopsy
6	Proven malignant by biopsy	Treatment

In this context, our contribution is a complete two-system CAD model, a detection CADe system and a description CADx system. The detection system aims to obtain an automated ROI of a tumor and its characteristics that will serve as input to the description system. We want to see that the results of the description system with handmade ROIs and automated ROIs are similar. Our description system will output not only every BI-RADS characteristic but also BI-RADS classification and malignancy estimation. We also want to prove that automatically obtained BI-RADS characteristics improve modern deep visual techniques for malignancy classification. Finally, this output will be returned in natural language, similar to what an expert radiologist would write in a medical report.

II. METHODS

A. Data Preparation

We have worked with three different public datasets that we named "BCD" ([21]), "B" ([22]), and "BUSIS" ([23]). BCD dataset contains 780 images classified as normal, benign, and malignant. Dataset B has 163 images with tumor type (cyst, fibroadenoma, etc.) and malignancy classification. Finally, BUSIS dataset has 562 images with no classification. Therefore, BUSIS dataset was mainly used for training the detection algorithm, while dataset B was preferred for the description system because the image quality was better, according to our radiologist expert. Our radiologist is a recognized expert in breast cancer with more than 30 years of experience in different hospitals. We found that BCD dataset contained duplicated images not only in the same class but between classes. We used Space-Invariant Feature Transform (SIFT) algorithm to detect zoomed or rotated copies of the same image [24]. We found that 150 images had almost exact copies at least once, 8 of them in a different class (Fig. 1). We eliminated 189 images. SIFT does not consider ultrasound images that are the same but moved in time, so these images were not eliminated from the database, but we did not use them for description or detection. Lastly, ground truth images do not consider all nodules, especially simple cysts. This dataset has been cited 205 times in google scholar and is typically used for the first step in algorithm development. We also applied SIFT for the other two datasets,

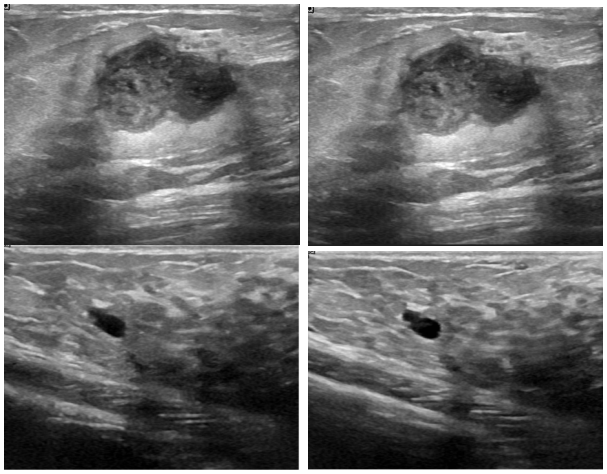

Fig. 1. First two images in the upper part are classified as copies with SIFT, one belongs to the benign class and the other to the malignant. Last two images are the same nodule moved in time and are not classified as copies.

finding two images in B and eight in BUSIS with a copy. For these reasons, we carefully selected the images used for detection and description, and also to avoid the Clever Hans phenomenon [25], discarding images with extra information such as color maps or ROIs. Clever Hans phenomenon refers to the algorithm learning patterns that may give better results in validation but do not represent the reality of the problem. For example, BCD dataset contains images with squares where the nodule is. Using them as training and validation for the detection system may give better results since the algorithm will learn that these squares are ROIs.

We obtained a total of 360 images with ROI labels based on ground truth and 139 with tumor descriptions from an expert in breast cancer. For these 139 images, we also took their malignancy classification.

B. Architecture

Our architecture has two differentiable systems, detection and description.

1) Detection system: The detection system uses YOLO algorithm that takes as input the ultrasound image and outputs the ROI for every tumor [26]. YOLO is a fully convolutional detection algorithm and, hence, can take different shaped inputs. Consequently, the obtained ROI has its original shape and can be introduced to the description system with no changes. YOLO is known for being fast, and its fifth version, Yolov5, runs at 50 frames per second, allowing a real-time detection system. It can also produce more than one ROI per image, detecting multiple nodules at once for later description.

Before introducing YOLO's output image to the descriptor, we used zero-padding to fill the ROIs to 450×450. The image was only resized when its width or height was higher than this value. This process was mainly done because ROIs have wide sizes (Table II), and resizing to a specific value would mean information loss. It is worth mentioning that applying zero-padding to standardize image size does not affect the CNNs training procedure and has no negative impact on time performance [27].

TABLE II
ROI'S SHAPES STATISTICS

Statistic	Width	Height	Ratio
Min	52	56	0.65
Max	620	436	2.72
Mean	224	183	1.26
STD	104	81	0.36

2) Description system: The description system takes as input a nodule extracted by YOLO and outputs the description in natural language with the BI-RADS characteristics, BI-RADS classification, and malignancy classification (learned from the databases). This system can be divided into three different parts. In Fig. 2, we can see the structure of the architecture we will describe below, with an example ROI of a fibroadenoma.

The first part inputs the image extracted from YOLO and outputs the BI-RADS descriptors and the malignancy classification. We tried two different autoencoder models. These models consist of an encoder that inputs the image and obtains abstract features, and a decoder that inputs these features and returns the model's output [28], in this case, the BI-RADS characteristics. The differences between the two models we propose lay in the decoder part, where we have an attention-based Long Short-Term Memory (LSTM) and an attention-based feedforward neural network. Both models have a convolutional layer with max-pooling and VGG16 as an encoder, with an output size of $14 \times 14 \times 512$ after the pooling layer, generating a 196×512 feature space. The encoder corresponds to the first part of Fig. 2.

- **LSTM decoder**: The first decoder model is an LSTM network with an attention mechanism [29]. Attention mechanisms were first developed for NLP autoencoder problems such as language translation, so the encoder's first words information was not lost when decoding to another language. This problem was solved by inputting a weighted average of the encoder's every hidden state value in each step of the decoder [30]. This mechanism quickly became popular in computer vision problems with excellent results [31], [32]. In our case, the attention mechanism takes the previous hidden state of the LSTM and the feature space as input and outputs a weighted average of the feature space named context [33]. This type of network has great results and is popular in image captioning problems. The context for iteration t is calculated as follows:

$$context_t = \sum_{i=1}^{196} f_i * a_{i,t} \tag{1}$$

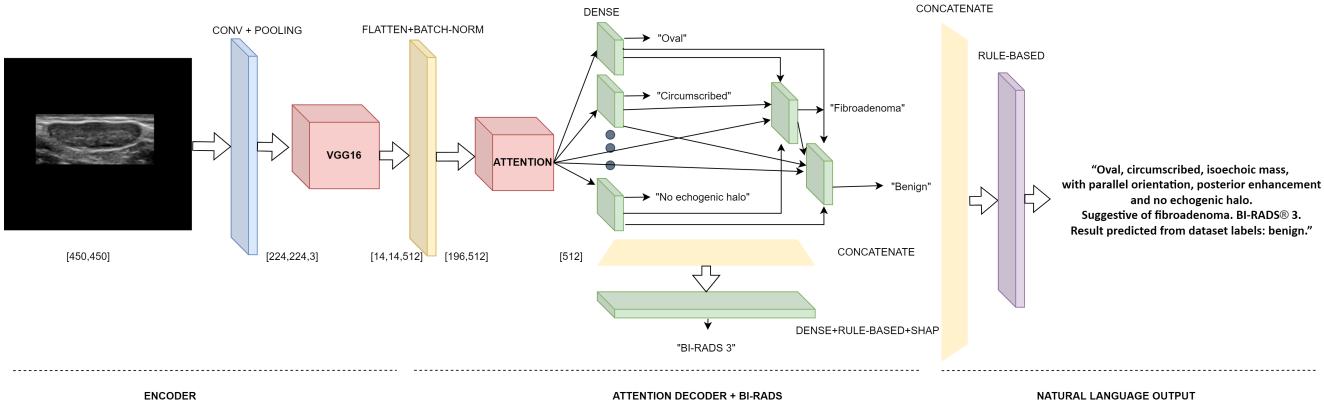

Fig. 2. Description system graphic for Attention decoder model. The encoder inputs the ROI with zero-padding. It passes through one convolutional layer with max-pooling and enters VGG16. This network gives the feature-space that is normalized and passed to the Attention decoder. The decoder returns the characteristics: “oval”, “circumscribed”, “isoechoic”, “parallel”, “posterior enhancement”, “no halo”, “fibroadenoma”, and “benign”. All of them, expect “benign”, are passed to the BI-RADS feedforward network, which outputs BI-RADS 3. Finally, the rule based model gives the natural language output.

where

$$a_{i,t} = \frac{exp(tanh(\mathbf{V}f_i + \mathbf{W}h_{t-1}))}{exp(\sum_{j=1}^{196} tanh(\mathbf{V}f_j + \mathbf{W}h_{t-1}))} \quad (2)$$

, $\mathbf{V}$ and $\mathbf{W}$ are weight matrices, and h_{t-1} is the hidden state for iteration $t-1$. This means that we first input each feature space vector into a perceptron with a hyperbolic tangent (tanh) activation function, which returns the importance of each vector ranging from -1 to 1. We then concatenate all of them into a softmax activation function that gives an ordered output of these importances ranging from 0 to 1, adding up to a total of 1, so we can calculate the weighted average of the feature space. These weights can also represent where the model has paid more attention to the image. Other more complex attention mechanisms such as two-layers and gated attention were tested during experimentation but with no improvements [34]. The context and the previously generated BI-RADS descriptors are the inputs of the LSTM, which outputs the following characteristic. For this network, we used a hidden state dimension of 256.

We used one rule to improve the system’s orientation output and overwrite the autoencoder’s result; if the nodule is round, there is no orientation because it is equally parallel and anti-parallel. Finally, since the model was trained with some nodules that were not classified as benign or malignant (extracted from the BUSIS dataset), it was possible that the LSTM decoder did not give a malignancy classification. Therefore, the BI-RADS output of the model was taken to calculate malignancy, where BI-RADS 4A or fewer means benign and 4B or more malignant.

- **Attention decoder**: The second decoder normalizes the 196×512 feature space with batch normalization, in contrast with the LSTM decoder, and converts it into a 512 dimensional space using the attention algorithm described above. In this case, the attention only inputs the image feature space. The 512 dimensional context is the input to six dense layers that output the shape, margin, orientation, echogenicity, posterior features, and halo. These layers have the same size as the number of variables in each descriptor and a softmax activation function as output, except for orientation and posterior that have sigmoid. This is because a nodule can be neither parallel nor anti-parallel; for example, round tumors, and posterior features are sometimes unclear, our expert did not always give a description. Therefore, our model only provides these two features when the result exceeds a threshold of 0.3. The same rule for orientation used in the LSTM decoder was incorporated into this system.

Once obtained the output of the layers of the six BI-RADS descriptors, the dense layers’ outputs and the 512 dimensional context were taken as input for the suggestivity dense layer. This layer has a softmax activation function and an extra label for no classification. This process was done because the suggestivity of a tumor depends on its BI-RADS characteristics. In contrast with the posterior characteristic, the extra no classification label was preferred to a sigmoid activation since suggestivity only makes sense in certain scenarios; 55 of the 139 described nodules did not have this characteristic. Sigmoid activation function resulted in more suggestivity outputs than needed, no matter the threshold. Therefore, if the model output was this last extra label, no suggestivity was chosen. Finally, all characteristics’ dense layers and the context were the input for malignancy classification dense layer of size two and softmax activation function. This structure can easily be visualized in Fig. 2, where the Attention decoder inputs the feature space and outputs the fibroadenoma BI-RADS characteristics and its

malignancy estimation.

The second part is a rule-based model and a feedforward neural network that inputs the BI-RADS descriptors and outputs the nodule final BI-RADS classification. The rule-based model is used when the nodule is classified as a simple cyst, complex cyst, or is spiculated, BI-RADS 2, BI-RADS 4A, and BI-RADS 5, respectively. The neural network has one hidden layer of size 12 to consider the connections between different characteristics. In Fig. 2, the feedforward neural network was used to output the final classification, BI-RADS 3. This means that the patient would have to return for another inspection in six months.

Finally, once we have extracted all the information from the nodule, a rule-based model gives the output in natural language as a radiologist would do in a medical report.

We applied SHapley Additive exPlanation (SHAP) to the BI-RADS classification feedforward network to obtain a more explainable model [35]. SHAP is a game theory-based algorithm that measures the importance of each variable in the input of a model to make a specific prediction. If each variable plays a player's role, and the prediction is the payout, SHAP estimates how to distribute that payout between players fairly. This decision is made not only considering their individual influence but how they act collectively.

C. Experiments

We have divided our experiments between detection and description systems.

1) Detection system: The detection system experiment uses 280 images for training and 80 for the test. We used precision, recall, and Area Under the Precision-Recall Curve (AUPRC), also known as Average Precision (AP), to evaluate our model. Precision and recall values vary depending on the threshold chosen to classify an object as positive; usually, the value is 0.5. The larger the threshold, the bigger the precision and the smaller the recall, and vice versa. It is possible to create a graphic with the axes being the recall and the precision, and the function their values when varying the threshold. AUPRC metric measures the area under this function.

YOLOv5 customized data augmentation was used. This strategy consists of random scaling, color space augmentations, and mosaic data loader (joining more than one image in one). When training, we used an established dimension for ultrasound images of 480×480; YOLOv5 version automatically applies zero-padding to maintain the image height-width ratio. This is crucial for some characteristics, such as form and orientation.

Finally, the test experiments for the description system indirectly show if YOLO could capture nodules' outside characteristics.

TABLE III
COHEN'S KAPPA EXPLANATION

Value	< 0	0 - 0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8 - 1
Agreement	Poor	Slight	Fair	Moderate	Substantial	Almost perfect

2) Description system: For the description algorithm, we have done three repetitions of 6-fold CV with 114 images and a test with 25 images to verify that no error was made. For CV, the ROI was hand chosen, while we took YOLO's output for the test. Random rotation with a range of $0.05 * 2\pi$ radians and horizontal flip were used as data augmentation techniques. The slight rotation is due to not altering tumor orientation. With this configuration, we have carried out three different experiments.

- **Database not related**: We made a database not related comparison between expert intracorrelation and intercorrelation values, Samsung's model, and our Attention model. We obtained expert intracorrelation and intercorrelation values from the papers presented in Section I, with a weighted average over the number of nodules [36]–[38]. To evaluate and compare the results, we used Cohen's kappa metric, the most used metric for expert agreement comparison in medicine. Table III shows the meaning of the results for this metric. To obtain Cohen's kappa for our model, we compared the results with the descriptions given by our expert.

- **Database related**: We did another experiment to compare how different model structures adjust to our database. Using Cohen's kappa metric, we obtained CV and test results for all BI-RADS descriptors with VGG16, Attention, and LSTM models.

- **Malignancy classification**: Finally, we created a new model based on the Attention model, with only the malignancy dense layer that inputs the context and outputs the classification. We named this model "Class model" and used it to compare classification results with our two proposed models to see if learning BI-RADS descriptors among classification improved the results. We also compared these results with our radiologist expert. The malignancy classification for our expert was obtained from the final BI-RADS classification as we did when the LSTM model did not output it. If the expert ranked the tumor as BI-RADS 4A or lower, it was classified as benign; otherwise, as malignant. We used accuracy, recall, precision, and F1 metrics in this experiment.

III. Results

Regarding the detection system, YOLO model obtained a precision of 0.965, recall of 0.95, and AP of 0.97. Fig. 3 shows how YOLO could also capture the part of the image where we find the necessary BI-RADS descriptors, such as the posterior feature.

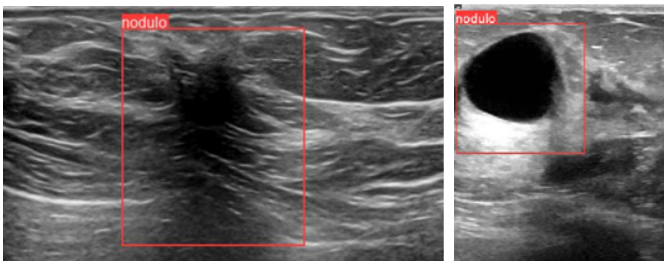

Fig. 3. YOLO detection test results for a malignant tumor with posterior shadowing and a cyst with posterior enhancement. Only the part of the image with the detected nodule is shown.

TABLE IV
DATABASE NOT RELATED EXPERIMENT

Characteristic	Intercorrelation	Intracorrelation	Samsung	CV	Test
Shape	0.48	0.68	0.64	0.45	**0.71**
Margin	0.34	0.59	0.3	0.46	**0.83**
Orientation	0.6	**0.76**	0.61	0.53	0.57
Echogenicity	0.35	**0.72**	0.34	0.45	0.4
Posterior	0.5	0.69	0.29	0.46	**0.75**
Halo	0.58	**0.72**	0.26	0.64	0.63
Suggestivity	-	-	-	**0.68**	0.52
BI-RADS	0.47	**0.75**	-	0.47	0.46
MEAN	0.47	**0.7**	0.41	0.49	0.62

The metric used is Cohen's kappa. Mean of model characteristics where obtained not considering suggestivity results, for better comparison with expert correlation.

In Table IV, we can see that test and CV means for database not related experiment were better than those obtained by Samsung's software and are between expert intercorrelation and intracorrelation values. For the halo characteristic, the other works also considered if the nodule had an abrupt interface or not. Samsung's software had better shape and orientation values than ours in CV, but, in the test, only in orientation. For most descriptors, the CV agreement is between expert intercorrelation and intracorrelation, except for shape, orientation, and posterior, which are lower. On the contrary, in the test, our model performed better in shape, margin, and posterior than expert intracorrelation and lower than intercorrelation again in orientation and BI-RADS.

Table V shows that in the database related experiment, Attention model and LSTM model had better results than VGG16 for every descriptor in CV, and only in echogenicity VGG16 was better than LSTM in test. The mean was better for Attention in both CV and test, achieving substantial agreement in the last, and only in margin did LSTM surpass the others.

Finally, Table VI shows how our models' performance was lower in tumor classification than our radiologist expert in the CV experiment but similar in test. Our models would have captured as malignant a nodule that our expert would have classified as BI-RADS 3 and would not have been biopsied. In contrast, the classification model had worse accuracy, precision, and F1 results than our model but better recall than the expert in CV. In test, the three models obtained the same results. As we can see, our expert and our two models have

TABLE V
DATABASE RELATED EXPERIMENT

	CV			Test		
Characteristic	VGG16	LSTM Model	Attention Model	VGG16	LSTM Model	Attention Model
Shape	0.38	0.42	**0.45**	0.45	**0.72**	0.71
Margin	0.44	**0.53**	0.46	0.75	**0.92**	0.83
Orientation	0.49	**0.54**	0.53	0.37	**0.57**	**0.57**
Echogenicity	0.21	0.37	**0.45**	0.19	0.17	**0.4**
Posterior	0.41	**0.46**	**0.46**	0.48	0.68	**0.75**
Halo	0.59	0.63	**0.64**	0.52	**0.63**	**0.63**
Suggestivity	0.44	0.51	**0.68**	0.25	0.44	**0.52**
MEAN	0.42	0.49	**0.52**	0.43	0.59	**0.63**

The metric used is Cohen's kappa. BI-RADS final classification wasn't considered because it is obtained using the same network.

TABLE VI
CLASSIFICATION RESULTS EXPERIMENT

	CV				Test	
Metric	Expert	LSTM Model	Attention Model	Class Model	Expert	All Models
Accuracy	0.86	**0.83**	**0.83**	0.77	0.88	0.88
Recall	0.68	0.64	0.64	**0.71**	0.75	**0.83**
Precision	0.93	**0.84**	0.83	0.67	**1**	0.91
F1	0.78	**0.73**	0.72	0.69	0.86	**0.87**

higher precision than recall. Precision is an essential metric since biopsy requires effort, money, and, most importantly, it is an invasive technique that may cause pain, bleeding, hemorrhaging, and hematomas [39].

In Fig. 4, the SHAP values of the example in Fig. 2 are shown. In red, we have the characteristics that have most influenced the prediction, while in blue, we have the opposite. We can see two different predictions, BI-RADS 3, the real output, and BI-RADS 5. The main reasons to classify the nodule as BI-RADS 3 are that it was suggestive of fibroadenoma, not a simple cyst, not anechoic, and was circumscribed. The reasons not to classify it as BI-RADS 5 were that it had no halo, no posterior shadowing, was not hypoechoic, and had parallel orientation. One interesting thing is that not having an indistinct margin was positive to choose BI-RADS 5. This can be because all our nodules classified as BI-RADS 5 were spiculated.

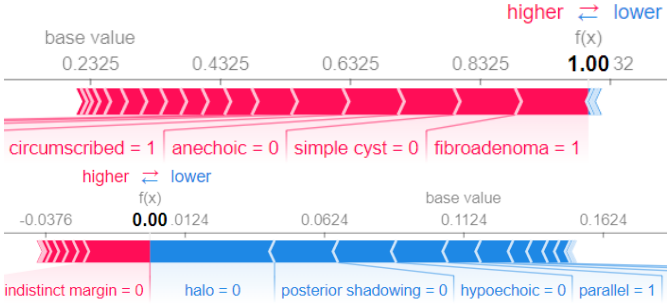

Fig. 4. SHAP output for the example of Fig. 4. First image explains why the output is BI-RADS 3 and the second one why it is not BI-RADS 5.

IV. Discussion

We found that a complete CAD for breast cancer ultrasound is possible and can be used in real-time video due to YOLO's detection fastness and efficiency. As we can see, in test results from the experiments on detection and description systems, YOLO can not only detect tumors but also capture their outside characteristics, the halo, and posterior features. Therefore, YOLO can be integrated with an image captioning system, like the one we propose, to describe nodules with BI-RADS descriptors. For the description system, unlike previous works, we have used for the test not handmade ROIs but the outputs of an earlier algorithm, with great results. We have followed every step of an expert during breast cancer diagnosis to obtain a model that helps and shortens their work as much as possible.

We have carried out three experiments for our description system during this process. With only 114 images for CV and 25 for the test, our models achieved an accordance level with our expert that is between expert intercorrelation and intracorrelation values, and have better results than Samsung's intercorrelation test. Also, we show that two different modern visual techniques trained with BI-RADS descriptors obtain better or equal results than the same model trained only for tumor classification. As we can see in CV for this malignancy experiment, our two models have a similar proportion of recall-precision to our expert, while the Class model obtained completely different results. This might mean that training the model with expert descriptions among classification has resulted in a model that classifies more like the expert. For this reason, we can conclude that the description given by the algorithm provides a good explanation of the final estimation of the tumor malignancy. This also implies that if an apparently benign tumor described as BI-RADS 2 or BI-RADS 3 is classified as malignant, it is worth checking it since the description boosts benign classification, meaning the visual part of the algorithm has detected something malignant. Finally, we incorporated the SHAP algorithm in our model for physicians to understand the BI-RADS final classification result. The output of this algorithm is visually attractive and easily explainable, which may increase expert trustworthiness in our model.

Our main limitation is the size of our database and the lack of description of nodule characteristics from more than one expert. Nevertheless, we have obtained significant results with our two-system-based model, and like in previous works, we have compared our description system with VGG16, outperforming it with a big difference [20]. Finally, having better test results than in CV experiments shows that our model has room to grow.

V. Conclusions

We have presented a complete CAD that guides and shortens the work of an expert in every step of tumor diagnosis, explaining different choices. First, the expert has a real-time detection mechanism for breast ultrasound video to aid in this phase. Then, once an interesting nodule is found, the expert can generate a description of it using the BI-RADS system, radiologists' universal language for tumor description. This description also outputs why the model has classified the nodule as a certain BI-RADS and can generate an explanation of why not to choose the others. At last, a final malignancy estimation is also included that can help doctors detect apparently benign cancers. Experimental results show that this system is applicable in real scenarios, assisting physicians in decision-making.

Future work includes adding a segmentation algorithm to the input of our description system that aims to improve tumor shape, margin, and orientation results. Based on recent works, we will also develop different Attention models for not correlated BI-RADS descriptors.

References

[1] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] V. McCormack *et al.*, "Breast cancer survival and survival gap apportionment in sub-Saharan Africa (ABC-DO): a prospective cohort study," *The Lancet Global health*, vol. 8, no. 9, pp. e1203–e1212, 2020.

[3] C. M. Ronckers, C. A. Erdmann, and C. E. Land, "Radiation and breast cancer: a review of current evidence," *Breast Cancer Research*, vol. 7, no. 1, pp. 1–12, 2004.

[4] H. Qi and N. A. Diakides, "Thermal infrared imaging in early breast cancer detection-a survey of recent research," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, vol. 2. IEEE, 2003, pp. 1109–1112.

[5] S. T. Kakileti, G. Manjunath, H. Madhu, and H. V. Ramprakash, "Advances in breast thermography," in *IntechOpen*, 2017, p. 91.

[6] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review," *Clinical imaging*, vol. 37, no. 3, pp. 420–426, 2013.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[9] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[13] "Data protection in EU," https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_es, European Comission, 2020.

[14] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Medical Image Analysis*, p. 102470, 2022.

[15] Q. Huang, Y. Chen, L. Liu, D. Tao, and X. Li, "On combining biclustering mining and AdaBoost for breast tumor classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 4, pp. 728–738, 2019.

[16] Q. Huang, B. Hu, and F. Zhang, "Evolutionary optimized fuzzy reasoning with mined diagnostic patterns for classification of breast tumors in ultrasound," *Information Sciences*, vol. 502, pp. 525–536, 2019.

[17] W. K. Moon, C.-M. Lo, J. M. Chang, C.-S. Huang, J.-H. Chen, and R.-F. Chang, "Quantitative ultrasound analysis for classification of BI-RADS category 3 breast masses," *Journal of digital imaging*, vol. 26, no. 6, pp. 1091–1098, 2013.

[18] J. Shan, S. K. Alam, B. Garra, Y. Zhang, and T. Ahmed, "Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods," *Ultrasound in medicine & biology*, vol. 42, no. 4, pp. 980–988, 2016.
[19] K. Kim, M. K. Song, E.-K. Kim, and J. H. Yoon, "Clinical application of S-Detect to breast masses on ultrasonography: a study evaluating the diagnostic performance and agreement with a dedicated breast radiologist," *Ultrasonography*, vol. 36, no. 1, p. 3, 2017.
[20] Q. Huang and L. Ye, "multi-task/single-task joint learning of ultrasound BI-RADS features," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2021.
[21] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE transactions on biomedical engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.
[22] M. H. Yap *et al.*, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1218–1226, 2017.
[23] Y. Zhang *et al.*, "BUSIS: a benchmark for breast ultrasound image segmentation," in *Healthcare*, vol. 10. MDPI, 2022, p. 729.
[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
[25] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.
[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
[27] M. Hashemi, "Enlarging smaller images before inputting into convolutional neural network: zero-padding vs. interpolation," *Journal of Big Data*, vol. 6, no. 1, pp. 1–13, 2019.
[28] D. Bank, N. Koenigstein, and R. Giryes, "Autoencoders," *arXiv preprint arXiv:2003.05991*, 2020.
[29] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
[31] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.
[32] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.
[33] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.
[34] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.
[35] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.
[36] E. Lazarus, M. B. Mainiero, B. Schepps, S. L. Koelliker, and L. S. Livingston, "BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value," *Radiology*, vol. 239, no. 2, pp. 385–391, 2006.
[37] C. S. Park *et al.*, "Observer agreement using the ACR breast imaging reporting and data system (BI-RADS)-ultrasound, (2003)," *Korean journal of radiology*, vol. 8, no. 5, pp. 397–402, 2007.
[38] H.-J. Lee *et al.*, "observer variability of breast imaging reporting and data system (BI-RADS) for breast ultrasound," *European journal of radiology*, vol. 65, no. 2, pp. 293–298, 2008.
[39] H.-L. Park and J. Hong, "Vacuum-assisted breast biopsy for breast cancer," *Gland surgery*, vol. 3, no. 2, p. 120, 2014.

6

Conclusiones y Líneas Futuras

En este trabajo hemos presentado la primera arquitectura CAD completa en ecografía de mama que ayuda al médico tanto en la detección, CADe, como en el diagnóstico, CADx, empleando el mismo método de clasificación que los radiólogos. Este sistema puede ser empleado en tiempo real gracias a la velocidad de YOLO. Una vez se encuentre un nódulo del cual el médico requiera descripción, esta se dará en el lenguaje universal BI-RADS, que es empleado también para conocer la probabilidad de malignidad del tumor y los procedimientos a seguir. Además, el modelo obtiene una estimación final de su malignidad que puede ayudar a reducir los falsos negativos. Por último, hemos incluido un algoritmo de explicación como SHAP para clarificar en que descriptores BI-RADS se ha basado la red para dar su clasificación final.

Durante los experimentos hemos obtenido resultados notables para el algoritmo YOLO en detección de nódulos, lo que coincide con los resultados encontrados en la literatura. Sin embargo, también hemos demostrado que este algoritmo es capaz no solo de detectar los nódulos, sino, también, de obtener las alteraciones que este provoca en el tejido cercano y en la ecografía. Esto queda reflejado en el halo ecogénico y la característica posterior, que son de gran importancia a la hora de detectar tumores malignos. Esto se ha logrado empleando en el test del sistema descriptor, por primera vez, ROI obtenidas por YOLO y no por un experto, lo cual supone una ventaja notable para la detección del cáncer de mama, por ejemplo, frente al sistema comercial vendido por Samsung [44] o el algoritmo de clasificación BI-RADS descrito en el estado del arte [45].

Utilizando tan solo 139 imágenes para validación cruzada y test, nuestro sistema descriptor (*Attention decoder*) ha obtenido valores de concordancia con nuestra experta que están entre los de intercorrelación e intracorrelación de los expertos, además de superar los resultados del experimento que se hizo con el modelo de Samsung.

En la clasificación de nódulos según su malignidad hemos comparado nuestros dos sistemas descriptores con un modelo idéntico, pero tan solo entrenado con la clasificación. En este experimento hemos obtenido mejores resultados al entrenar los modelos también con las características BI-RADS, logrando además una proporción precisión-exhaustividad similar a la de nuestra experta. Esto indica que estos modelos han aprendido a razonar de forma más similar a la experta a través de los descriptores.

Durante el escrito de este trabajo continuamos con las reuniones con la experta para obtener una base de datos mayor. Cuando la base de datos era de menor tamaño, el

sistema con LSTM obtenía mejores resultados que el simple, pero, cuanto más crecía el número de nódulos con descripción, menor era la diferencia. En las últimas pruebas con la base de datos con la que se ha presentado el artículo, la red con capa densa ha obtenido mejores resultados. Esto puede deberse a que la LSTM, al tomar en cada iteración la palabra anterior, aprendía descripciones de nódulos típicas, como las de los quistes simples; por ejemplo, "nódulo ovalado, circunscrito, con orientación paralela y refuerzo posterior, sugestivo de quiste simple (BI-RADS 2)". A medida que el número de nódulos ha ido creciendo, la red densa con atención simple ha aprendido a generalizar mejor, ya que no considera la salida de la anterior característica. Además, cabe mencionar que la salida de la red densa se ha hecho más compleja en las últimas pruebas hasta llegar a la que presentamos en el artículo. Por lo tanto, seguimos explorando con ambos sistemas descriptores y modificando la arquitectura para obtener mejores resultados.

En trabajos futuros estudiaremos, también, entrenar la red por separado con los diferentes descriptores y emplear los filtros CLAHE y SRBF. Asimismo, probaremos con un nuevo filtro que será el resultado de un algoritmo de segmentación. Este filtro podría ayudar a obtener mejores resultados en las características que se basen en la morfología del nódulo: la forma, el margen y la orientación.

El sistema detector YOLO admite como entrada imágenes de diferente tamaño, pero nuestro sistema descriptor no. Esto lo hemos solucionado añadiendo *zero-padding* a las imágenes de entrada del sistema detector para no alterar las propiedades de los nódulos y no perder información. Sin embargo, gracias al algoritmo de atención, es posible programar este sistema para que adquiera esta propiedad. Así, todo el modelo en su conjunto no dependería del tamaño de la imagen de entrada y no sería necesario realizar ningún escalado ni *padding* para poder introducir las imágenes a estas redes. La red *encoder* del sistema descriptor que da la entrada al mecanismo de atención es una red completamente convolucional. Como hemos visto en el caso de YOLO, esta red podría admitir imágenes de diferente tamaño, provocando que su salida también sea distinta. Así, obtendríamos una salida $n \times n \times p$, donde n sería un valor que dependería del tamaño de entrada. Sin embargo, esto no es problema para el algoritmo de atención que fue pensado para el caso de las RNN, que admiten entradas de tamaño diferente. Sin importar el valor de n, el mecanismo de atención siempre devuelve el contexto de dimensión p. Esta es la codificación de la imagen que se emplea en el sistema *decoder*, por lo que sin importar el tamaño de la imagen, la entrada al decoder siempre sería p.

La descripción del nódulo en el sistema BI-RADS no es del todo objetiva y existe discrepancia entre expertos. Entrenar el modelo tan solo con una experta puede llevar

a un sistema demasiado personalizado y que no convenza a la mayoría, cuantos más expertos participen en las descripciones, más neutral será el modelo. Por ello, consultaremos a más expertos con el objetivo de que el sistema sea menos personalizado y más global. Asimismo, obtendríamos una gran base de datos para la explicación de la clasificación BI-RADS que podría ser estudiada mediante el algoritmo SHAP.

Buscamos que el modelo sea interactivo y que el experto pueda corregir y mejorar los resultados del sistema. Es más, este modelo podría aprender de dichos cambios y adaptarlo a las preferencias de cada usuario.

Por último, el objetivo final de esta investigación es poner en sistema a prueba en casos reales. En julio de 2022 hemos estado en Makeni, Sierra Leona y hemos recogido vídeos de ecografías de pacientes con cáncer de mama, grabados por el técnico local del Holy Spirit Hospital, mediante una sonda Philips Lumify lineal conectada a una tablet. Próximamente, emplearemos el modelo para analizarlas y, en el futuro, nuestro modelo se probará con esa misma sonda en ese hospital. Más adelante y si los resultados son los esperados, el modelo comenzará a emplearse en más hospitales.

Este trabajo se inscribe dentro del proyecto "Cribado coste-efectivo del cáncer de mama mediante mamografía, ecografía y termografía", financiado por el Ministerio de Ciencia e Innovación (ref. PID2019-110686RB-I00), en el cual soy investigador.

Bibliografía

[1] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] V. McCormack *et al.*, "Breast cancer survival and survival gap apportionment in sub-Saharan Africa (ABC-DO): a prospective cohort study," *The Lancet Global health*, vol. 8, no. 9, pp. e1203–e1212, 2020.

[3] C. M. Ronckers, C. A. Erdmann, and C. E. Land, "Radiation and breast cancer: a review of current evidence," *Breast Cancer Research*, vol. 7, no. 1, pp. 1–12, 2004.

[4] H. Qi and N. A. Diakides, "Thermal infrared imaging in early breast cancer detection-a survey of recent research," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No. 03CH37439)*, vol. 2. IEEE, 2003, pp. 1109–1112.

[5] S. T. Kakileti, G. Manjunath, H. Madhu, and H. V. Ramprakash, "Advances in breast thermography," in *IntechOpen*, 2017, p. 91.

[6] K. J. Taylor, C. Merritt, C. Piccoli, R. Schmidt, G. Rouse, B. Fornage, E. Rubin, D. Georgian-Smith, F. Winsberg, B. Goldberg, and E. Mendelson, "Ultrasound as a complement to mammography and breast examination to characterize breast masses," *Ultrasound in medicine & biology*, vol. 28, no. 1, pp. 19–26, 2002.

[7] A. Jalalian, S. B. Mashohor, H. R. Mahmud, M. I. B. Saripan, A. R. B. Ramli, and B. Karasfi, "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review," *Clinical imaging*, vol. 37, no. 3, pp. 420–426, 2013.

[8] L. Liberman and J. H. Menell, "Breast imaging reporting and data system (BI-RADS)," *Radiologic Clinics*, vol. 40, no. 3, pp. 409–430, 2002.

[9] E. Lazarus, M. B. Mainiero, B. Schepps, S. L. Koelliker, and L. S. Livingston, "BI-RADS lexicon for US and mammography: interobserver variability and positive predictive value," *Radiology*, vol. 239, no. 2, pp. 385–391, 2006.

[10] C. S. Park *et al.*, "Observer agreement using the ACR breast imaging reporting and data system (BI-RADS)-ultrasound, (2003)," *Korean journal of radiology*, vol. 8, no. 5, pp. 397–402, 2007.

[11] H.-J. Lee *et al.*, "observer variability of breast imaging reporting and data system (BI-RADS) for breast ultrasound," *European journal of radiology*, vol. 65, no. 2, pp. 293–298, 2008.

[12] P. Skaane, K. Engedal, and A. Skjennald, "Interobserver variation in the interpretation of breast imaging: comparison of mammography, ultrasonography, and both combined in the interpretation of palpable noncalcified breast masses," *Acta Radiologica*, vol. 38, no. 4, pp. 497–502, 1997.

[13] L. J. Grimm, A. L. Anderson, J. A. Baker, K. S. Johnson, R. Walsh, S. C. Yoon, and S. V. Ghate, "Interobserver variability between breast imagers using the fifth edition of the BI-RADS MRI lexicon," *American Journal of Roentgenology*, vol. 204, no. 5, pp. 1120–1124, 2015.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[16] *Directrices Éticas para una IA fiable*. Grupo Independiente de Expertos de Alto Nivel Sobre Inteligencia Artificial. Comisión Europea, 2019. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60423

[17] "Data protection in EU," https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_es, European Comission, 2020.

[18] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[21] S.-C. Lo, S.-L. Lou, J.-S. Lin, M. T. Freedman, M. V. Chien, and S. K. Mun, "Artificial convolution neural network techniques and applications for lung nodule detection," *IEEE transactions on medical imaging*, vol. 14, no. 4, pp. 711–718, 1995.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis," *Medical Image Analysis*, p. 102470, 2022.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[29] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[30] Z. Niu, G. Zhong, and H. Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[31] A. Sinha and J. Dolz, "Multi-scale self-guided attention for medical image segmentation," *IEEE journal of biomedical and health informatics*, vol. 25, no. 1, pp. 121–130, 2020.

[32] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[33] G. H. Aly, M. Marey, S. A. El-Sayed, and M. F. Tolba, "YOLO based breast masses detection and classification in full-field digital mammograms," *Computer methods and programs in biomedicine*, vol. 200, p. 105823, 2021.

[34] Z. Cao, L. Duan, G. Yang, T. Yue, and Q. Chen, "An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures," *BMC medical imaging*, vol. 19, no. 1, pp. 1–9, 2019.

[35] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," *arXiv preprint arXiv:1410.1090*, 2014.

[36] M. Alsharid, H. Sharma, L. Drukker, P. Chatelain, A. T. Papageorghiou, and J. A. Noble, "Captioning ultrasound images automatically," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 338–346.

[37] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[38] A. Singh, J. Krishna Raguru, G. Prasad, S. Chauhan, P. K. Tiwari, A. Zaguia, and M. A. Ullah, "Medical image captioning using optimized deep learning model," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[39] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "VQA-med: overview of the medical visual question answering task at ImageCLEF 2019." *CLEF (Working Notes)*, vol. 2, 2019.

[40] Q. Huang, Y. Chen, L. Liu, D. Tao, and X. Li, "On combining biclustering mining and AdaBoost for breast tumor classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 4, pp. 728–738, 2019.

[41] Q. Huang, B. Hu, and F. Zhang, "Evolutionary optimized fuzzy reasoning with mined diagnostic patterns for classification of breast tumors in ultrasound," *Information Sciences*, vol. 502, pp. 525–536, 2019.

[42] W. K. Moon, C.-M. Lo, J. M. Chang, C.-S. Huang, J.-H. Chen, and R.-F. Chang, "Quantitative ultrasound analysis for classification of BI-RADS category 3 breast masses," *Journal of digital imaging*, vol. 26, no. 6, pp. 1091–1098, 2013.

[43] J. Shan, S. K. Alam, B. Garra, Y. Zhang, and T. Ahmed, "Computer-aided diagnosis for breast ultrasound using computerized BI-RADS features and machine learning methods," *Ultrasound in medicine & biology*, vol. 42, no. 4, pp. 980–988, 2016.

[44] K. Kim, M. K. Song, E.-K. Kim, and J. H. Yoon, "Clinical application of S-Detect to breast masses on ultrasonography: a study evaluating the diagnostic performance and agreement with a dedicated breast radiologist," *Ultrasonography*, vol. 36, no. 1, p. 3, 2017.

[45] Q. Huang and L. Ye, "multi-task/single-task joint learning of ultrasound BI-RADS features," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 2021.

[46] F. Wang *et al.*, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[47] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?.^ex^plaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[48] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[49] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 38, no. 1, pp. 35–44, 2004.

[50] M. Ravikumar, P. Rachana, B. Shivaprasad, and D. Guru, "Enhancement of mammogram images using clahe and bilateral filter approaches," in *Cybernetics, Cognition and Machine Learning Applications*. Springer, 2021, pp. 261–271.

[51] S. Balocco, C. Gatta, O. Pujol, J. Mauri, and P. Radeva, "SRBF: Speckle reducing bilateral filtering," *Ultrasound in medicine & biology*, vol. 36, no. 8, pp. 1353–1363, 2010.

[52] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE transactions on biomedical engineering*, vol. 63, no. 7, pp. 1455–1462, 2015.

[53] M. H. Yap *et al.*, "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE journal of biomedical and health informatics*, vol. 22, no. 4, pp. 1218–1226, 2017.

[54] Y. Zhang *et al.*, "BUSIS: a benchmark for breast ultrasound image segmentation," in *Healthcare*, vol. 10. MDPI, 2022, p. 729.