UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Aplicación de modelos de aprendizaje profundo basados en convoluciones y transformadores para la detección de masas en imágenes de mamografías.

Alejandro Javier Duque Aguilera

Ingeniería en Ciencias de la Computación

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
Ingeniero en Ciencias de la Computación

Quito, 15 de Diciembre de 2023

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Aplicación de modelos de aprendizaje profundo basados en convoluciones y transformadores para la detección de masas en imágenes de mamografías.

Alejandro Javier Duque Aguilera

Nombre del profesor, Título académico **Noel Pérez Pérez, Doctor en Informática**

Quito, 15 de Diciembre de 2023

© DERECHOS DE AUTOR

Nombres y apellidos: Alejandro Javier Duque Aguilera

Código: 00209216

Cédula de identidad: 1725863490

Lugar y fecha: Quito, 15 de Diciembre de 2023

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Este estudio se centra en la aplicación de arquitecturas Transformer para la detección de masas en imágenes mamográficas digitales extraídas de la base de datos InBreast. Se dedica especial atención a la evaluación de la arquitectura DETR (Detection Transformer) y su variante deformable. Se investiga el impacto de varios hiperparámetros de la arquitectura DETR en el rendimiento de estos modelos. Los resultados revelan que la versión deformable del DETR demuestra una mayor adaptabilidad y un rendimiento superior en comparación con el DETR original, logrando un mAP50 de 0.681 y un mAP50:95 de 0.405. Sin embargo, al compararse con arquitecturas convolucionales como YOLOv8, el DETR deformable muestra limitaciones especialmente en la detección de masas pequeñas. Por último, se destaca la importancia de generar modelos DETR preentrenados de diferentes tamaños para lograr una aplicabilidad más amplia en diversos dominios.

Palabras clave: Detección de Masas, Mamografías, InBreast, Transformers, DETR, YOLO, Aprendizaje Profundo

ABSTRACT

This research focuses on the application of Transformer architectures for mass detection in digital mammographic images extracted from the InBreast database. Special attention is devoted to the evaluation of the DETR (Detection Transformer) architecture and its deformable variant. The impact of various hyperparameters of the DETR architecture on the performance of these models is investigated. The results reveal that the deformable version of the DETR demonstrates greater adaptability and superior performance compared to the original DETR, achieving a mAP50 of 0.681 and a mAP50:95 of 0.405. However, when compared to convolutional architectures such as YOLOv8, deformable DETR shows limitations especially in small mass detection. Finally, the importance of generating pre-trained DETR models of different sizes is highlighted to achieve broader applicability in various domains.

Keywords: Mass Detection, Mammograms, InBreast, Transformers, DETR, YOLO, Deep Learning

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

El cáncer, caracterizado por el rápido crecimiento y división celular, conduce a la formación de tumores que representan un riesgo significativo debido a su capacidad para invadir el tejido circundante (National Institute of Health US, 2007). Entre los diversos tipos de cáncer, el de mama destaca como uno de los más comunes a nivel mundial, contribuyendo solo en 2020 a 685,000 muertes (Organización Mundial de la Salud, 2020). Un diagnóstico temprano se vuelve esencial para iniciar un tratamiento oportuno, aunque en el caso del cáncer de mama, este desafío se agrava al intentar identificar la región del tejido con lesiones a través de imágenes de mamografía. La variabilidad en las formas y la densidad del tejido circundante dificultan esta tarea (Pérez, 2015).

Los métodos tradicionales para detectar masas mamarias, como los autoexámenes mamarios (Roth et al., 2018; Huang et al., 2022), exámenes clínicos mamarios (Huang et al., 2022), mamografías (Bassett & Gold, 1987; Nover et al., 2009; Moore, 2001; Nounou et al., 2015; Nyström et al., 1993; Park & Ikeda, 2006), doble lectura de mamografías (Taylor-Phillips, & Stinton, 2020), ecografía (Moore, 2001; Park & Ikeda, 2006; Ozmen et al., 2015), imágenes por resonancia magnética (MRI) (Moore, 2001; Nounou et al., 2015; Park & Ikeda, 2006) y biopsias (Nyström et al., 1993), a menudo generan errores debidos al factor humano involucrado en cada enfoque. Por lo tanto, se vuelve crucial explorar alternativas de detección que no dependan del factor humano, sino que capitalicen el poder computacional del aprendizaje profundo.

En los últimos años, la detección de objetos en el ámbito médico ha confiado exclusivamente en arquitecturas convolucionales, según lo indicado por Ganatra et al. (2021). Esta tendencia se mantiene al abordar específicamente la detección de masas mamarias, como destaca el análisis realizado por Mahoro & Akhloufi (2022). Las arquitecturas convolucionales utilizadas para la detección de objetos se agrupan principalmente en dos categorías fundamentales: detectores de objetos con propuestas regionales de dos etapas, como Faster-RCNN (Ren et al., 2015), y detectores de objetos de una sola etapa, como YOLO (Redmon et al., 2016). La eficacia de estos enfoques ha sido respaldada por investigaciones como las de Agarwal (2020), Akselrod et al. (2016), Fan et al. (2019) y Peng et al. (2020), que se centraron

en el uso de Faster-RCNN, así como por Al-Masni et al. (2018), Aly et al. (2021) y Baccouche et al. (2021), quienes optaron por arquitecturas YOLO. Además, se están explorando arquitecturas específicamente diseñadas para la detección de masas mamarias, como BMassDNet desarrollado por Cao et al. (2021), RoI CNN propuesto por Bhatti et al. (2020), y la modificación PAA propuesta por Jiang et al. (2022).

A pesar de que las redes neuronales convolucionales han demostrado su efectividad para la detección de masas mamarias, los últimos años han visto el surgimiento de nuevas arquitecturas de aprendizaje profundo basadas en Transformers (Vaswani et al., 2017). Estas arquitecturas, conocidas por su éxito en procesamiento del lenguaje natural, han comenzado a adoptarse lentamente en problemas de visión artificial gracias al desarrollo del Vision Transformer (Dosovitskiy et al., 2020; Carion et al., 2020). En cuanto a la detección de masas mamarias, ya se pueden encontrar aplicaciones de Transformers que van desde la extracción de características (Betancourt, 2023) hasta la detección de objetos (Su et al., 2022). A pesar de estos avances, queda aún mucha investigación por realizar en este campo.

En 2020, investigadores de Facebook presentaron el Detection Transformer (DETR), una arquitectura basada en Transformers enfocada en la detección de objetos (Carion et al., 2020). Desde entonces, han surgido variantes como el Detection Transformer deformable de Zhu et al. (2020), o el DINO DETR propuesto por Zhang et al. (2022). Aunque la arquitectura DETR ha demostrado ser prometedora en bases de datos referenciales como COCO, hay una escasez notable de investigación relacionada con su aplicación específica en la detección de masas mamarias. Por lo tanto, proponemos desarrollar un método de aprendizaje profundo que maximice el rendimiento de la detección de masas mamarias en imágenes de mamografía utilizando la arquitectura DETR.

ESTADO DEL ARTE

La detección de objetos aborda el desafío de identificar elementos dentro de una imagen y asignarlos a categorías específicas. En el ámbito de la detección de masas mamarias, las arquitecturas convolucionales han destacado debido a su sólida capacidad de extracción de características y eficacia en el establecimiento de relaciones espaciales.

Diversos investigadores han empleado arquitecturas de propuestas regionales para abordar este propósito. Agarwal et al. utilizaron un modelo Faster R-CNN para detectar masas mamarias en la extensa base de datos de imágenes de mamografía OPTIMAM (OMI-DB), logrando un TPR de 0,93 con 0,78 falsos positivos por imagen (FPI). Este mismo modelo obtuvo un TPR de 0,85 a 0,1 FPI en el conjunto de datos de InBreast (Agarwal et al., 2020). Akselrod-Ballin et al. entrenaron un Faster R-CNN modificado en un conjunto de datos que comprendía 850 imágenes generadas por un sistema de datos e informes de imágenes mamarias (BI-RADS), resultando en una precisión media promedio (mAP) de 0,6 (Akselrod-Ballin et al., 2016). Fan et al. desarrollaron un sistema CAD (detección asistida por computadora) con Faster R-CNN entrenado en imágenes de tomo síntesis digital de mama (DBT) que logró un ROC AUC de 0,96 (Fan et al., 2019). Peng et al. integraron convoluciones deformables en el backbone de Faster R-CNN junto con un NAS-FPN para un sistema CAD que arrojó un TPR de 0,93 a 2,28 FPI en CBIS-DDSM y un TPR de 0,95 a 0,38 FPI en InBreast (Peng et al., 2020).

Otros autores han optado por arquitecturas de detección de objetos de una sola etapa. Al-masni et al. implementaron un sistema CAD basado en YOLO para la detección y clasificación simultánea de masas mamarias. El sistema se entrenó en 600 mamografías de la base de datos digital para mamografías de detección (DDSM) y lograron un AUC de 0,96 (Al-Masni et al., 2018). Por su parte, Aly et al. propusieron un detector de masa mamaria basado en YOLOv4, que alcanzó un mAP del 95,83 % cuando se entrenó en el conjunto de datos InBreast, superando en un 8% al mAP de un modelo YOLOv3 entrenado en las mismas condiciones (Aly et al., 2021). Baccouche entrenó un modelo YOLOv5 en 1.471 imágenes de mamografía del Instituto Nacional del Cáncer de Tailandia y, tras la optimización de hiperparámetros, logró un mAP de 0,91 (Baccouche et al., 2021).

Aunque muchos investigadores utilizan modelos de aprendizaje profundo bien establecidos, algunos prefieren implementar sus propias arquitecturas adaptadas al funcionamiento interno específico de la detección de masas mamarias. Un ejemplo es la propuesta de Cao et al., la Red Masiva de Detección Mamaria (BMassDNet), que consiste en una red convolucional basada en redes piramidales de características (FPN). BMassDNet logró 0,495 falsos positivos con un recall de 0,930 cuando se entrenó en InBreast y 0,599 falsos positivos con recall de 0,943 cuando se entrenó en DDSM (Cao et al., 2021). Bhatti et al. también optaron por una arquitectura personalizada que utilizaba CNN y FPN basados en RoI, logrando un mAP de 0,84 cuando se entrenó en el conjunto de datos DDSM (Bhatti et al., 2020). Jiang et al. propusieron una red PAA modificada, un detector de objetos sin anclaje de una sola etapa, y la combinaron con un clasificador de ROI. Esta arquitectura compuesta logró un mAP de 0,651 cuando se entrenó en un conjunto de datos de 3835 imágenes integradas de conjuntos de datos públicos CBIS-DDSM, InBreast y MIAS (Jiang et al., 2022).

A pesar de la prevalencia de modelos de aprendizaje profundo convolucionales, las arquitecturas de Transformers están empezando a ser exploradas en este contexto. La investigación en detección de masas mamarias con Transformers es aún limitada, con excepciones notables como el trabajo de Su et al., quienes propusieron YOLO-LOGO, un modelo de segmentación YOLO basado en Transformers para la detección de masas mamarias que obtuvo un mAP del 65,0 % en el conjunto de datos CBIS-DDSM (Su et al., 2022). Otra contribución destacada es la de Betancourt et al., quienes incorporaron una columna vertebral transformadora Swin en un detector de masas mamarias, logrando una TPR del 75,7 % a 0,1 FPI en el conjunto de datos OMI-DB (Betancourt, 2023).

A pesar de estos avances, la exploración de arquitecturas de Transformers aplicadas a la detección de masas mamarias ha pasado por alto en gran medida la arquitectura DETR, diseñada específicamente para la detección de objetos. Este proyecto tiene como objetivo llenar este vacío explorando las aplicaciones potenciales de DETR y su variante deformable para la detección de masas mamarias.

METODOLOGÍA EXPERIMENTAL

Base de datos

El éxito de todas las implementaciones de aprendizaje profundo se encuentra intrínsecamente ligado a la disponibilidad de datos adecuados para entrenar el modelo. Sin embargo, en el ámbito de las aplicaciones médicas, la obtención de datos suele ser un desafío significativo debido a las estrictas regulaciones destinadas a salvaguardar la privacidad de los pacientes. Este estudio utiliza el conjunto de datos de acceso público InBreast, que se compone de 410 imágenes de mamografías digitales de campo completo obtenidas en un centro de mama ubicado en un hospital universitario de Oporto. Estas imágenes abarcan diversos tipos de lesiones, como masas, calcificaciones, asimetrías y distorsiones. Es relevante destacar que los cuadros delimitadores de masas mamarias se fundamentan en anotaciones reales realizadas por expertos, como se detalla en el trabajo de Moreira et al. (2012).

Transformers

Las arquitecturas de Transformers han ganado considerable atención en los últimos años. Originalmente concebidos por Vaswani et al. para el análisis de texto, los Transformers han demostrado una destacada capacidad para aprender relaciones dentro de secuencias, atribuible en gran medida a sus mecanismos de atención (Vaswani et al., 2017). En 2020, Facebook introdujo DETR, marcando un hito al aplicar Transformers a tareas de detección de objetos (Carion et al., 2020). El concepto central de la arquitectura DETR implica la fusión de un bloque codificador-decodificador con una red convolucional. En este diseño, los canales de salida de la capa convolucional forman una secuencia que sirve como entrada del codificador. Este último utiliza mecanismos de atención para establecer relaciones entre los canales, mientras que el decodificador asocia un número fijo de consultas (del inglés “queries”) con la información de la imagen extraída por el codificador, generando posteriormente predicciones de clase y cuadro delimitador para cada consulta.

En años recientes, han surgido variantes y mejoras de DETR, orientadas principalmente a reducir el tiempo de entrenamiento. Una mejora notable es el DETR deformable, propuesto por Zhu et al. (2020) e inspirado en las convoluciones deformables ideadas por primera vez por Dai et al. (2017). El DETR deformable acelera la convergencia hasta diez veces en

comparación con el DETR base, mejorando simultáneamente el rendimiento en la detección de objetos pequeños. Esta mejora se logra mediante módulos de atención deformables que restringen la cantidad de píxeles atendidos por cada píxel. El DETR deformable se presenta como una excelente opción de arquitectura para la detección de masas mamarias, pues ha demostrado ser más efectivo en la detección de objetos pequeños que el DETR original (Zhu et al., 2020).

Tanto en las arquitecturas DETR como en las DETR deformables, varios hiperparámetros clave desempeñan un papel crucial en la comprensión de su funcionamiento. Estos incluyen el backbone convolucional, la dimensión del modelo, la cantidad de consultas y la cantidad de capas codificadoras-decodificadoras. El backbone convolucional comprende todo el bloque de capas convolucionales que inicialmente procesan la imagen, extrayendo las características secuenciales que alimentarán al codificador. La dimensión del modelo corresponde al tamaño de los vectores dentro de la secuencia de entrada del codificador, que a su vez consiste en una representación aplanada de los canales de salida de la capa final del backbone convolucional. Una característica crucial de las arquitecturas de Transformers es que la dimensión de entrada de una capa coincide con su dimensión de salida. Esto no solo simplifica la implementación de estas arquitecturas, sino que también facilita la creación de conexiones residuales dentro de las capas, mitigando el problema del gradiente que desaparece y acelerando la convergencia. El parámetro 'número de consultas' se refiere a varios factores esenciales: la longitud de la secuencia de entrada del codificador, el número de consultas en el decodificador y el número máximo de detecciones posibles en una imagen. Por último, las capas de codificación y decodificación se pueden apilar en cualquier cantidad deseada, gracias a la propiedad mencionada anteriormente de coincidencia de dimensiones, lo que simplifica enormemente la escalabilidad y adaptabilidad de la arquitectura.

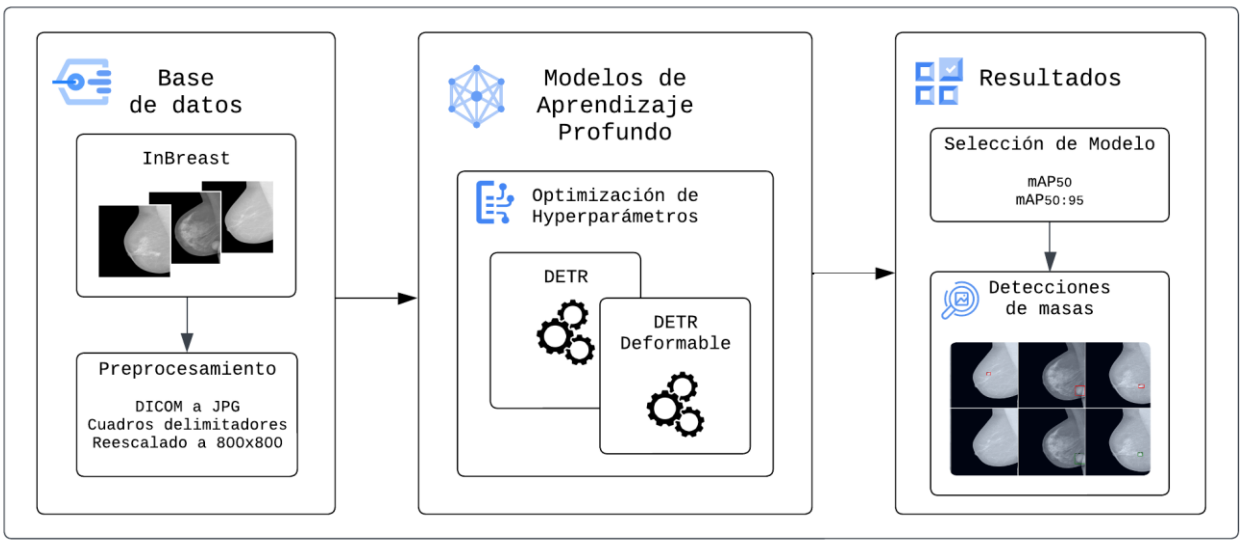
Método Propuesto

Nuestro enfoque propuesto se ilustra en la Figura 1. Iniciamos el proceso con el preprocesamiento del conjunto de datos para su preparación antes de ingresarlos a los modelos de aprendizaje profundo. Posteriormente, llevamos a cabo el entrenamiento del DETR y del DETR deformable, optimizando los hiperparámetros mediante validación cruzada. A continuación, seleccionamos el modelo que presenta el rendimiento más destacado, evaluado

mediante mAP50:95. Finalmente, establecemos el umbral óptimo de detección para el detector de objetos seleccionado.

El código desarrollado para este proyecto está disponible en GitHub (https://github.com/grimloc-aduque/Breast-Masses-Detection-using-Detection-Transformer-Architectures). La infraestructura empleada para ejecutar todos los experimentos incluye un contenedor Docker que se ejecuta en un servidor equipado con una tarjeta gráfica NVIDIA A100 de 80 GB proporcionada por la USFQ. La imagen Docker se encuentra alojada en Docker Hub (https://hub.docker.com/repository/docker/grimloc13/detr-cuda/general).

Figura 1. Flujo de trabajo del método propuesto.



Generación de Datasets

Las imágenes originales de mamografía en formato DICOM, obtenidas de InBreast, fueron convertidas al formato JPG mediante el uso de la biblioteca de Python dicom2jpg. Posteriormente, se extrajeron los cuadros delimitadores correspondientes a las masas mamarias a partir de las máscaras de la región de interés (ROI). Luego, todas las imágenes fueron redimensionadas a 800x800 píxeles, que representa el tamaño mínimo admitido por la arquitectura DETR. Para asegurar la alineación de las coordenadas de los cuadros delimitadores con el nuevo tamaño de imagen, se utilizó la biblioteca de Python albumentations. Tanto la imagen como las anotaciones delimitadoras se registraron en un archivo annotations.json en formato COCO. Tras completar esta etapa de preprocesamiento, el

conjunto de datos se dividió en subconjuntos de entrenamiento y prueba mediante una proporción del 90% para entrenamiento y 10% para prueba.

Métricas de Validación

El modelo DETR genera un número de predicciones de cuadros delimitadores equivalente a la cantidad de consultas. Cada cuadro predicho cuenta con una probabilidad asociada. La evaluación de la calidad de estas predicciones implica considerar dos factores fundamentales: el umbral de detección y el umbral de intersección sobre unión (IoU).

Estableciendo un umbral de detección específico, solo se consideran las detecciones con una probabilidad superior a dicho umbral. El emparejamiento entre estos cuadros delimitadores predichos y el conjunto de cuadros delimitadores reales se efectúa mediante el algoritmo de coincidencia húngaro. Para clasificar un cuadro delimitador como verdadero positivo (TP) o falso negativo (FN), se evalúa la superposición entre el cuadro delimitador predicho y el real utilizando la métrica de IoU. Un TP se registra si el IoU supera el umbral especificado; en caso contrario, se considera un FN. Las detecciones que no se corresponden con ningún cuadro delimitador real se categorizan como falsos positivos (FP). Al utilizar el recuento total de TP, FP y FN en todas las imágenes de validación/prueba, podemos calcular la precisión y el recall del modelo. Variando el umbral de detección de 0 a 1, es posible trazar una curva PR con cada par de datos de precisión y recall.

La precisión promedio (AP) para un umbral de IoU dado corresponde al área bajo la curva PR en este umbral de IoU. La precisión promedio media (mAP) corresponde al promedio de AP para todas las clases. En nuestro caso, AP y mAP son equivalentes, ya que solo tenemos una clase. Con este contexto sobre AP y mAP, podemos definir las dos métricas principales utilizadas para evaluar el rendimiento de nuestros modelos, que son mAP50 y mAP50:95. mAP50 corresponde al área bajo la curva PR en un umbral de IoU de 0,5, promediado para todas las clases. Mientras tanto, mAP50:95 es el área promedio bajo la curva PR para umbrales de IoU que varían de 0,5 a 0,95 con un paso de 0,05, promediado para todas las clases.

Configuración Experimental

Se emplearon las implementaciones DETR y DETR deformable de la biblioteca Transformers de Python, inicializándolos con pesos preentrenados obtenidos de HuggingFaces, específicamente de los checkpoints facebook/detr-resnet-50 y SenseTime/deformable-detr.

Estos checkpoints fueron originalmente entrenados en 118,000 imágenes del conjunto de datos de detección de objetos COCO 2017. Para aprovechar este preentrenamiento, establecimos una tasa de aprendizaje relativamente baja de 1e-5 para la backbone convolucional y 1e-4 para el bloque codificador-decodificador y los pesos del cabezal de detección. Optamos por utilizar AdamW como optimizador de pesos, que implementa una regularización de caída de peso desacoplada (Loshchilov & Hutter, 2017).

Se aplicó un aumento de datos sobre la marcha para proporcionar al modelo un conjunto de imágenes más diverso, contribuyendo así a mejorar sus habilidades de generalización. Este aumento de datos incluyó operaciones como reflexiones horizontales y verticales, eliminación de píxeles y transformaciones afines que implicaban escalado, traslación y rotación. La evaluación de las métricas de validación mAP50 y mAP50:95 se realizó con un umbral de detección de 0.001, utilizando la clase CocoEvaluator de la biblioteca coco-eval de Python.

DETR y Deformable DETR se entrenaron en InBreast mediante validación cruzada de 10 pliegues. Cada modelo se sometió a 200 épocas de entrenamiento con un tamaño de lote de 16. El entrenamiento se llevó a cabo utilizando la biblioteca Lightning de Python, incorporando detención temprana que supervisaba la pérdida de validación y detenía el entrenamiento después de 30 épocas sin mejoras. En cada pliegue, se calcularon mAP50 y mAP50:95 en el conjunto de validación correspondiente, y posteriormente se promediaron los resultados.

En cuanto a la optimización de hiperparámetros, se exploraron diferentes backbones (resnet10, resnet26, resnet50 de la biblioteca timm de Python) y diversas configuraciones de Transformers, incluyendo variaciones en el número de capas de codificador-decodificador (2, 4, 6, 8, 10), el número de consultas (50, 75, 100, 125, 150) y la dimensión del modelo (128, 192, 256, 320, 384). Debido a la cantidad considerable de hiperparámetros considerados, no fue posible realizar una exploración de búsqueda en grilla. En cambio, fijamos la configuración DETR original, que consta del backbone convolucional resnet50, 6 capas de codificador-decodificador, 100 consultas y una dimensión de modelo de 256. A partir de esta configuración base, exploramos el espacio de hiperparámetros, variando uno por uno para comprender el efecto de cada componente en el rendimiento del modelo.

ANÁLISIS DE RESULTADOS

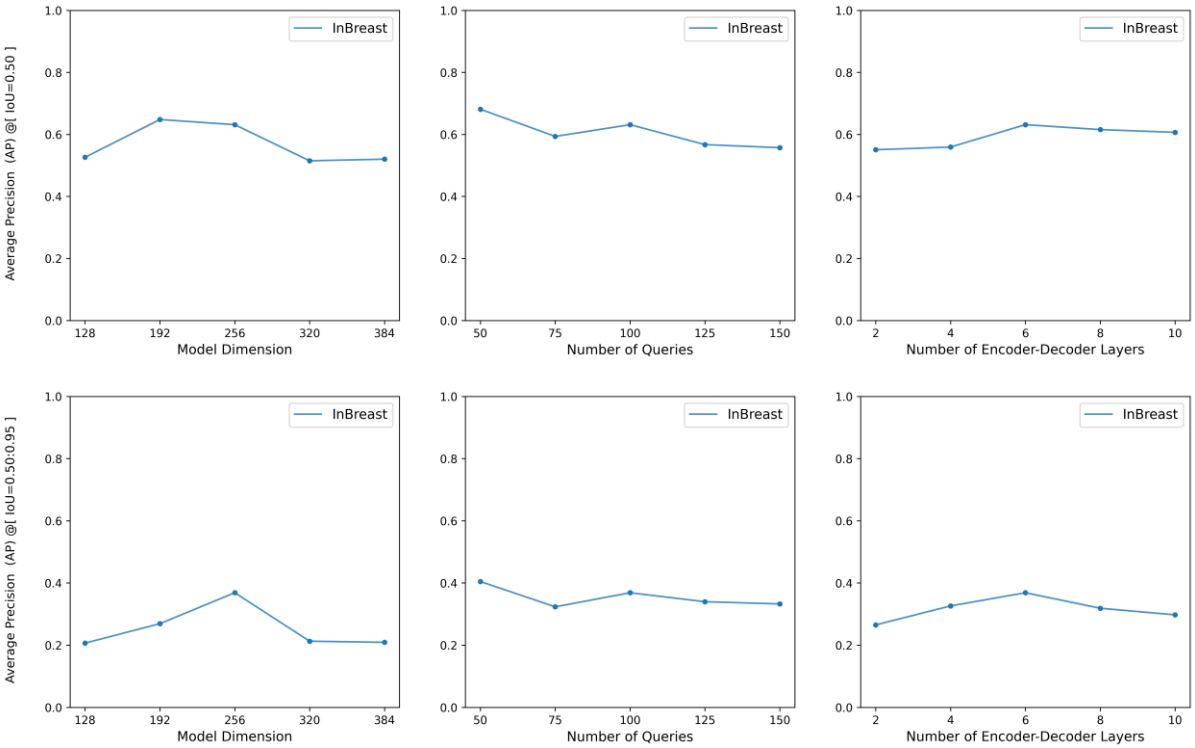
Optimización de Hiperparámetros

Durante la optimización de hiperparámetros, se realizaron descubrimientos significativos sobre las capacidades de aprendizaje de las arquitecturas DETR. Tanto el DETR como el DETR deformable demostraron ser sensibles a cambios en la backbone convolucional, revelando que no lograron aprender a detectar masas mamarias cuando se cambiaron de un resnet50 a un resnet26 o resnet10. Este fenómeno no pareció ser atribuible a la tasa de aprendizaje, ya que se exploraron diversas tasas, oscilando entre 1e-5 y 1e-3. Una hipótesis plausible es que la magnitud del conjunto de datos no resultó suficiente para reentrenar el extractor de características en el backbone del DETR. A diferencia del resnet50 original, estos backbones no fueron preentrenados específicamente para generar características secuenciales que un bloque codificador pueda procesar con eficacia.

En lo que respecta a las modificaciones en los bloques codificador-decodificador, se observaron resultados notoriamente diferentes entre DETR y DETR deformable. La arquitectura DETR original no logró entrenarse exitosamente al variar cualquier hiperparámetro. Modificar el número de capas, el número de consultas o la dimensión del modelo requería la reinicialización de pesos en secciones específicas de este bloque. Como señalaron Zhu et al. (Zhu et al., 2020), la arquitectura DETR original enfrenta desafíos para entrenar desde cero, especialmente en un conjunto de datos pequeño como InBreast.

En contraste, el DETR deformable demostró una capacidad de entrenamiento exitosa desde cero en estos bloques codificador-decodificador. Los gráficos de la Figura 2 ilustran las variaciones tanto en mAP50 como en mAP50:95 en relación con los cambios en los valores de los hiperparámetros para la arquitectura DETR deformable.

Figura 2. Optimización de métricas para DETR deformable en InBreast.



La configuración DETR deformable original, con 6 capas de codificador-decodificador, 100 consultas y una dimensión del modelo de 256, logra un mAP50 de 0.631 y un mAP50:95 de 0.369. Al variar la dimensión del modelo, el mAP50 aumenta ligeramente a 0.648 en la dimensión 192, indicando una mejora marginal con respecto a la configuración original. Mientras tanto, el mAP50:95 no presenta mejoras significativas. Al modificar el número de consultas, tanto mAP50 como mAP50:95 alcanzan su punto máximo en 50 consultas, registrando valores de 0.681 y 0.405, respectivamente. Aunque esto representa una mejora, no es demasiado sustancial respecto a sus valores originales. Al variar el número de capas de codificador-decodificador, ni mAP50 ni mAP50:95 muestran mejoras notables.

Según estos resultados, el hiperparámetro que tiene el mayor impacto en el rendimiento del modelo es el número de consultas. Para evaluar que estas diferencias son estadísticamente significativas se realiza una prueba estadística sobre el valor del mAP50:95 tomando como pivote el modelo de 50 consultas que es el de mejor rendimiento promedio y comparándolo frente a todos los otros modelos. Debido al tamaño de muestra se selecciona la prueba estadística t-Student pareada y de una sola cola. Se parte de la hipótesis nula de que el modelo

de 50 consultas no presenta un mAP50:95 superior a otro modelo, o lo que es lo mismo que la diferencia entre su rendimiento promedio y el de otro modelo es menor o igual que 0 (H_0: $\mu \leq 0$). Se tabula el valor p de cada una de esta prueba estadística en la tabla 1.

Tabla 1. Test estadístico t-Student sobre el mAP50:95 de modelos DETR Deformables

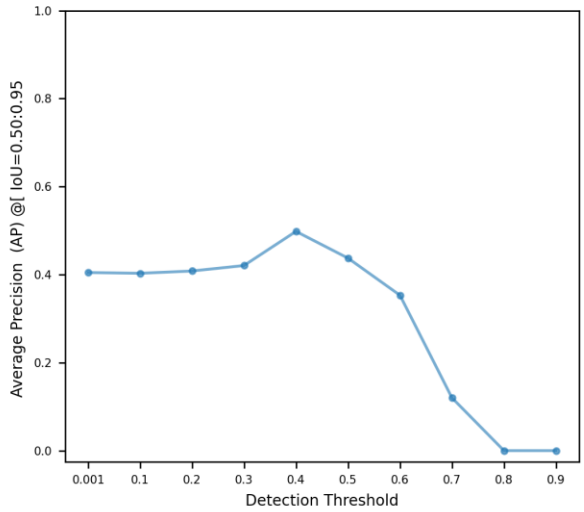
Dimensión del modelo	**Número de consultas**	**Número de capas codificador-decodificador**	**μ (mAP50:95)**	**Valor-p (mAP50:95)**
Arquitectura Original				
256	100	6	0.368	1.84 e-04
Optimización de dimensión de modelo				
128	100	6	0.207	6.72 e-10
192	100	6	0.269	1.26 e-09
320	100	6	0.213	3.37 e-10
384	100	6	0.209	4.00 e-10
Optimización de consultas				
256	50	6	0.405	Pivote
256	75	6	0.323	4.96 e-07
256	125	6	0.340	6.07 e-07
256	150	6	0.333	1.89 e-07
Optimización de capas codificador-decodificador				
256	100	2	0.265	4.38 e-07
256	100	4	0.326	3.63 e-06
256	100	8	0.319	2.85 e-07
256	100	10	0.297	1.28 e-07

Incluso considerando un valor de significancia $\alpha = 0.001$ se halla que el modelo de 50 consultas es estadísticamente superior que el resto de los modelos. Específicamente, la reducción a la mitad del número de consultas proporciona una mejora de 0.05 en mAP50 con respecto a la arquitectura original, sugiriendo que esta tendencia podría mantenerse incluso con un número aún menor de consultas que los explorados en este trabajo. Tanto la arquitectura DETR como la DETR deformable fueron diseñadas originalmente en el conjunto de datos COCO, que tiene 91 categorías y un promedio de 5 objetos por imagen. Para esa densidad de objetos, utilizar 100 o más consultas parece ser la estrategia correcta. No obstante, para tareas más específicas, como la detección de masas mamarias, con una sola clase y un promedio de aproximadamente un objeto por imagen, se sugiere reducir significativamente el número de consultas.

Detecciones de masas

Tras realizar el análisis de la optimización de hiperparámetros, con un enfoque particular en mAP50:95, se determinó que la mejor configuración para la arquitectura DETR es el DETR deformable, con una backbone resnet50, 50 consultas, 6 capas de codificador-decodificador y un tamaño de dimensión de 256. Para definir su umbral de detección óptimo, se evaluó el mAP50:95 en diversos umbrales, variando de 0 a 0.9 con incrementos de 0.1. El gráfico de optimización del umbral de detección se muestra en la Figura 3, donde el pico corresponde a un umbral de detección de 0.4, alcanzando un máximo de 0.498 en mAP50:95.

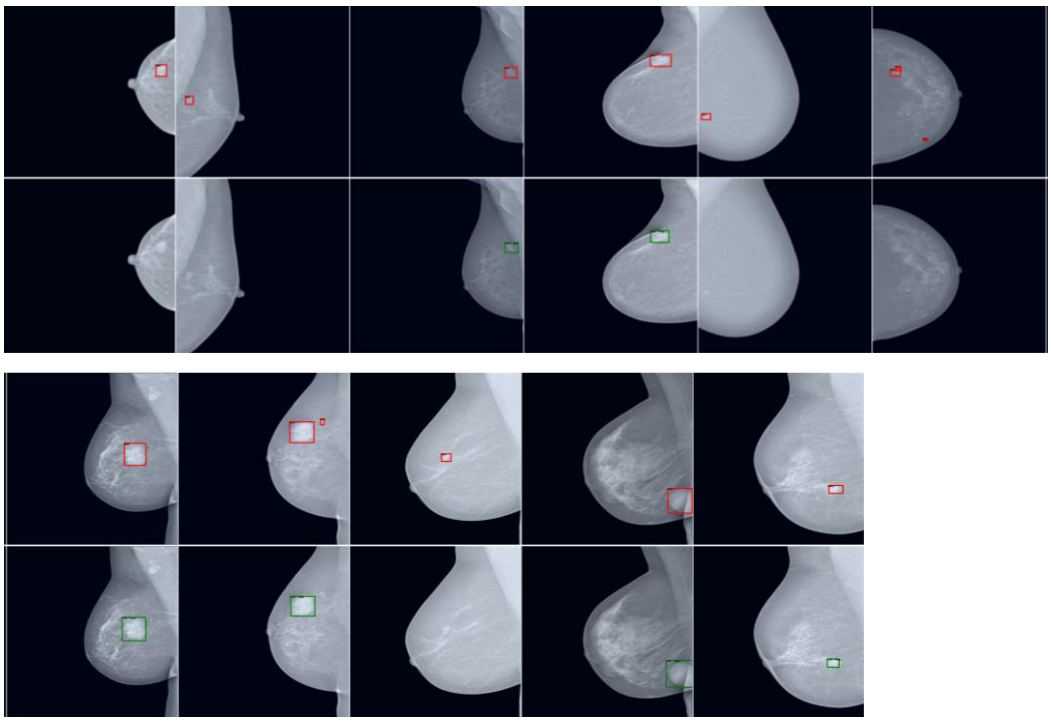
Figura 3. Optimización del umbral de detección del mejor modelo.



La Figura 4 presenta una comparación entre los cuadros delimitadores de masas mamarias reales y las detecciones generadas por el mejor modelo DETR deformable con el umbral de detección óptimo de 0.4. En la fila superior, se encuentran los datos reales resaltados en rojo, mientras que, en la fila inferior, se muestran las detecciones en verde. Aquí, las imágenes que logran detecciones de objetos están cerca de sus verdades fundamentales. Sin embargo, este escenario se observa solo en 6 de las 11 imágenes que conforman el conjunto de prueba. En contraste, las otras 5 imágenes del conjunto de prueba no obtienen ninguna detección. Se destaca que las masas no detectadas comparten una característica común: su tamaño reducido. Se podría considerar la posibilidad de reducir el umbral de detección para mejorar la identificación de estas masas pequeñas. Al reducir el umbral a 0.3, se logra detectar

algunas de las masas mamarias previamente no identificadas. No obstante, este ajuste tiene un efecto contraproducente al introducir numerosas detecciones de falsos positivos.

Figura 4. Comparación entre las masas mamarias reales y las detecciones.



Comparación con Modelos Convolucionales

Con el objetivo de lograr una interpretación más sólida de los resultados obtenidos con el modelo DETR, se consideró crucial establecer un punto de comparación. En este sentido, se optó por entrenar un modelo convolucional YOLOv8 en el conjunto de datos InBreast. La elección de YOLOv8 se basó en su sobresaliente rendimiento en la detección de objetos, previamente validado en la base de datos COCO 2017 (Ultralytics, 2021). La implementación de YOLOv8 de la biblioteca Ultralytics en Python ofrece versiones diferentes, distinguidas por el tamaño del modelo, de las cuales se evaluaron tres: nano (Yolov8n), pequeña (Yolov8s) y mediana (Yolov8m).

Cada modelo YOLOv8 se inicializó con pesos preentrenados y se entrenó utilizando el mismo esquema de validación cruzada de 10 pliegues empleado con el DETR. Los modelos fueron entrenados durante 200 épocas con un tamaño de lote de 16, implementando una estrategia de detención temprana después de 30 épocas. La Tabla 1 presenta los resultados de mAP50 y mAP50:95 para las tres versiones de YOLOv8, junto con las métricas

correspondientes a los mejores DETR y DETR deformable. Entre las versiones de YOLOv8, la nano destacó con un mAP50:95 de 0.573, un rendimiento sorprendente dada su modesta cantidad de 3.2 millones de parámetros, superando significativamente a un modelo de 25.9 millones de la misma familia.

Tabla 2. Comparación de modelos en base a mAP50 y mAP50:95

Modelo		**Métricas**		**Valor-p (mAP50:95)**
Arquitectura	**# Parámetros (M)**	**μ (mAP50)**	**μ (mAP50:95)**	
YOLOv8				
Nano: Yolov8n	3.2	0.826	0.573	Pivote
Pequeño: Yolov8s	11.2	0.836	0.540	1.56 e-04
Mediano: Yolov8m	25.9	0.829	0.522	3.45 e-06
DETR				
Backbone = resnet50, Dimensión = 256, Consultas = 100, Capas = 6	40.0	0.281	0.232	2.04 e-11
DETR Deformable				
Backbone = resnet50, Dimensión = 256, Consultas = 50, Capas = 6	39.5	0.681	0.405	7.31 e-10

La tabla 2 presenta una comparación de los modelos YOLOv8 con los DETR. Resulta evidente que cada versión de YOLOv8 supera al DETR deformable en ambas métricas con un margen significativo. Esta disparidad puede atribuirse al tamaño de los modelos, ya que el DETR deformable cuenta con 39.5 millones de parámetros, superando en diez veces la cantidad de parámetros de YOLOv8 nano que fue el de mejor rendimiento. Se aplica la misma prueba estadística explicada en la tabla 1, pero este caso usando de pivote el modelo YOLOv8 nano, y se encuentra que esta diferencia en mAP50:95 si es estadísticamente significativa.

El tamaño del modelo DETR plantea un problema, dado que si la tarea de detección se puede llevar a cabo con tan solo 3.2 millones de parámetros, el exceso de parámetros en DETR se vuelve redundante. En nuestros experimentos, intentamos abordar esto mediante la reducción del tamaño del DETR original, reemplazando el backbone resnet50 de 23.5 millones de parámetros por resnet10 de 4.9 millones de parámetros o resnet26 de 13.9 millones de parámetros. Sin embargo, estos intentos resultaron infructuosos debido a las dificultades de aprendizaje inherentes al DETR.

Es evidente que el empleo del modelo DETR según su diseño original no parece ser la opción más acertada para la detección de masas en imágenes mamográficas, dado que un

modelo convolucional más ligero, como el YOLOv8 nano, exhibe un rendimiento superior. No obstante, si se pretende explorar arquitecturas de Transformers como el DETR en el futuro, será esencial contar con versiones preentrenadas en diversos tamaños, siguiendo la variedad ofrecida por YOLOv8.

CONCLUSIONES

En esta investigación se evaluaron las arquitecturas basadas en Transformers, centrándose particularmente en el Detection Transformer y su variante deformable, para la detección de masas en imágenes de mamografía mediante ajustes y optimizaciones de hiperparámetros utilizando el conjunto de datos InBreast, que abarca mamografías digitales.

Entre los hallazgos clave, se destaca la sensibilidad de las arquitecturas DETR y DETR deformable ante la utilización de nuevos backbones convolucionales. Se observó que ningún DETR logró aprender a detectar masas mamarias al cambiar el backbone resnet50 original por uno de menor tamaño. Este desafío de aprendizaje fue recurrente en el DETR, ya que la arquitectura tampoco pudo reentrenarse efectivamente frente a cualquier reinicialización de pesos en su bloque Transformer. En contraste, el DETR deformable demostró la capacidad de reentrenar su bloque codificador-decodificador incluso frente a reinicializaciones de pesos.

En el análisis de la optimización de hiperparámetros, se encontró que reducir el número de consultas es esencial en aplicaciones de dominio donde la densidad promedio de objetos por imagen es baja. La mejor configuración identificada para el DETR incluye un backbone resnet50, 50 consultas, 6 capas de codificador-decodificador y una dimensión del modelo de 256. Este modelo deformable, con un mAP50:95 de 0.405, logra detectar con éxito masas medianas y grandes, aunque no es tan eficiente en la detección de masas pequeñas.

A pesar de la mejora en la capacidad de aprendizaje del DETR deformable en comparación con el DETR original, no logró superar el rendimiento de YOLOv8. La versión nano de YOLOv8 alcanzó un mAP50:95 de 0.573, destacando su superioridad en este contexto. Estos resultados indican avances significativos en la detección de masas mamarias con arquitecturas de Transformers, pero subrayan la importancia continua de la investigación y desarrollo en este campo. Se destaca la necesidad crucial de desarrollar modelos DETR

preentrenados de diversos tamaños para mejorar la capacidad de adaptación a las complejidades y variaciones presentes en datos médicos, señalando una dirección fundamental para futuros desarrollos en este ámbito.

Por último, se plantea la incógnita sobre la generalización de los resultados a otras bases de datos de imágenes mamográficas, como DDSM. DDSM dispone de una mayor cantidad de imágenes, lo que podría resultar beneficioso para el proceso de aprendizaje de los modelos DETR. Este aspecto podría ser objeto de análisis en futuras investigaciones.

REFERENCIAS BIBLIOGRÁFICAS

Agarwal, R., Diaz, O., Yap, M. H., Lladó, X., & Marti, R. (2020). Deep learning for mass detection in full field digital mammograms. Computers in biology and medicine, 121, 103774.

Akselrod-Ballin, A., Karlinsky, L., Alpert, S., Hasoul, S., Ben-Ari, R., & Barkan, E. (2016). A region based convolutional network for tumor detection and classification in breast mammography. In Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 2 (pp. 197-205). Springer International Publishing.

Al-Masni, M. A., Al-Antari, M. A., Park, J. M., Gi, G., Kim, T. Y., Rivera, P., ... & Kim, T. S. (2018). Simultaneous detection and classification of breast masses in digital mammograms via a deep learning YOLO-based CAD system. Computer methods and programs in biomedicine, 157, 85-94.

Aly, G. H., Marey, M., El-Sayed, S. A., & Tolba, M. F. (2021). YOLO based breast masses detection and classification in full-field digital mammograms. Computer methods and programs in biomedicine, 200, 105823.

Baccouche, A., Garcia-Zapirain, B., Olea, C. C., & Elmaghraby, A. S. (2021). Breast Lesions Detection and Classification via YOLO-Based Fusion Models. Computers, Materials & Continua, 69(1).

Bassett, L. W., & Gold, R. H. (1987). Breast cancer detection: mammography and other methods in breast imaging.

Betancourt Tarifa, A. S., Marrocco, C., Molinara, M., Tortorella, F., & Bria, A. (2023). Transformer-based mass detection in digital mammograms. Journal of Ambient Intelligence and Humanized Computing, 14(3), 2723-2737.

Bhatti, H. M. A., Li, J., Siddeeq, S., Rehman, A., & Manzoor, A. (2020, December). Multi-detection and segmentation of breast lesions based on mask rcnn-fpn. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2698-2704). IEEE.

Cao, H., Pu, S., Tan, W., & Tong, J. (2021). Breast mass detection in digital mammography based on anchor-free architecture. Computer Methods and Programs in Biomedicine, 205, 106033.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Cham: Springer International Publishing.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 764-773).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fan, M., Li, Y., Zheng, S., Peng, W., Tang, W., & Li, L. (2019). Computer-aided detection of mass in digital breast tomosynthesis using a faster region-based convolutional neural network. Methods, 166, 103-111.

Ganatra, N. (2021, March). A comprehensive study of applying object detection methods for medical image analysis. In 2021 8th international conference on computing for sustainable global development (INDIACom) (pp. 821-826). IEEE.

Huang, N., Chen, L., He, J., & Nguyen, Q. D. (2022). The efficacy of clinical breast exams and breast self-exams in detecting malignancy or positive ultrasound findings. Cureus, 14(2).

Jiang, J., Peng, J., Hu, C., Jian, W., Wang, X., & Liu, W. (2022). Breast cancer detection and classification in mammogram using a three-stage deep learning framework based on PAA algorithm. Artificial Intelligence in Medicine, 134, 102419.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Mahoro, E., & Akhloufi, M. A. (2022). Applying Deep Learning for Breast Cancer Detection in Radiology. Current Oncology, 29(11), 8767-8793.

Moore, S. K. (2001). Better breast cancer detection. Ieee Spectrum, 38(5), 50-54.

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. Academic radiology, 19(2), 236-248.

National Institute of Health US. (2007). Understanding cancer - NIH curriculum supplement series - NCBI bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK20362/

Nounou, M. I., ElAmrawy, F., Ahmed, N., Abdelraouf, K., Goda, S., & Syed-Sha-Qhattal, H. (2015). Breast cancer: conventional diagnosis and treatment modalities and recent

patents and technologies. Breast cancer: basic and clinical research, 9, BCBCR-S29420.

Nover, A. B., Jagtap, S., Anjum, W., Yegingil, H., Shih, W. Y., Shih, W. H., & Brooks, A. D. (2009). Modern breast cancer detection: a technological review. Journal of Biomedical Imaging, 2009, 1-14.

Nyström, L., Wall, S., Rutqvist, L. E., Lindgren, A., Lindqvist, M., Rydén, S., ... & Larsson, L. G. (1993). Breast cancer screening with mammography: overview of Swedish randomised trials. The Lancet, 341(8851), 973-978.

Ozmen, N., Dapp, R., Zapf, M., Gemmeke, H., Ruiter, N. V., & van Dongen, K. W. (2015). Comparing different ultrasound imaging methods for breast cancer detection. IEEE transactions on ultrasonics, ferroelectrics, and frequency control, 62(4), 637-646.

Park, J. M., & Ikeda, D. M. (2006). Promising techniques for breast cancer detection, diagnosis, and staging using non-ionizing radiation imaging techniques. Physica Medica, 21, 7-10.

Peng, J., Bao, C., Hu, C., Wang, X., Jian, W., & Liu, W. (2020). Automated mammographic mass detection using deformable convolution and multiscale features. Medical & biological engineering & computing, 58, 1405-1417.

Pérez, N. P. (2015). Improving Variable Selection and Mammography-based Machine Learning Classifiers for Breast Cancer CADx (Doctoral dissertation, Universidade do Porto (Portugal)).

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28.

Roth, M. Y., Elmore, J. G., Yi-Frazier, J. P., Reisch, L. M., Oster, N. V., & Miglioretti, D. L. (2011). Self-detection remains a key method of breast cancer detection for US women. Journal of women's health, 20(8), 1135-1139.

Su, Y., Liu, Q., Xie, W., & Hu, P. (2022). YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms. Computer Methods and Programs in Biomedicine, 221, 106903.

Taylor-Phillips, S., & Stinton, C. (2020). Double reading in breast cancer screening: considerations for policy-making. The British journal of radiology, 93(1106), 20190610.

Ultralytics. (2021). *GitHub - Ultralytics/Ultralytics: NEW - YOLOV8 in PyTorch*. https://github.com/ultralytics/ultralytics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

World Health Organization. (2020). Breast cancer. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/Breast-Cancer

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., ... & Shum, H. Y. (2022). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.