Redes Siamesas vs Transformers

y el

Uso de Redes GANS

Autoras:

Johanna Angulo

Jenny Aguilar

Fecha: 7 de abril de 2025

Índice

1. Redes siamesas en imágenes térmicas de mama

Las redes neuronales siamesas son modelos de *deep learning* compuestos por dos subredes idénticas (con pesos compartidos) entrenadas para medir la similitud entre dos imágenes de entrada[1]. En el contexto de termografía mamaria, este enfoque explota la simetría bilateral natural de la temperatura en ambas mamas: en mujeres sanas la distribución térmica es similar en ambos senos, mientras que la presencia de un tumor suele romper dicha simetría (generando mayor calor en la zona afectada). Un ejemplo reciente es el trabajo de Dey *et al*. (2024), que introdujo una red siamesa para detección de anomalías mamarias en termogramas comparando la imagen térmica del seno izquierdo vs. derecho de la misma paciente. Esta red siamesa aprendió a distinguir pares similares (mamas sin lesión, de pacientes sanas) de pares diferentes (mamas con asimetría térmica por tumor) y, con un conjunto de datos público relativamente pequeño, logró una exactitud del 81% en la clasificación de pacientes con anomalías. Este enfoque representa la primera aplicación documentada de redes siamesas en termografía mamaria, aprovechando explícitamente el análisis bilateral para la detección de cáncer de mama. Cabe destacar que las redes siamesas son consideradas métodos de aprendizaje de pocos ejemplos (*few-shot learning*), lo cual las hace adecuadas para conjuntos de datos reducidos (algo común en imágenes médicas) al aprender a reconocer diferencias relevantes sin requerir miles de muestras por clase.

Además de la detección binaria (normal vs. anormal) mediante asimetría, las redes siamesas se han explorado en termografía para clasificación multiclase con datos limitados. Por ejemplo, Ervural y Ceylan (2021) aplicaron un modelo siamés a imágenes térmicas de neonatos para predecir distintas patologías, combinando estrategias de aumento de datos para paliar la escasez de muestras[2]. En su estudio de 44 neonatos (4 categorías diagnósticas), reportan que la mejor precisión de clasificación alcanzó ~94%, y aproximadamente 89% al evaluar en pacientes completamente nuevos no vistos en entrenamiento. Esto demuestra que, incluso con muy pocos datos, una red profunda apoyada en técnicas siamesas y aumentos sintéticos puede lograr desempeños altos en termografía médica. No obstante, se observa que el rendimiento cae al enfrentar casos verdaderamente nuevos (89% vs. 94%), subrayando la importancia de generalizar más allá de los datos entrenados. En general, las redes siamesas han demostrado utilidad en tareas médicas donde se dispone de pares de imágenes correlacionadas. Por ejemplo, en mamografía digital (otra modalidad de imagen mamaria), Bai *et al*. (2022) desarrollaron una red siamesa que compara la mamografía actual de una paciente con

---

[1] Dey, A., & Rajan, S. (2024). Thermography-based Breast Abnormality Detection using Siamese Network. *CMBES Proceedings*, *46*. Retrieved from https://proceedings.cmbes.ca/index.php/proceedings/article/view/1186

[2] Ervural, S., & Ceylan, M. (2022). Classification of neonatal diseases with limited thermal image data. *Multimedia Tools and Applications, 81*, 1–29. https://doi.org/10.1007/s11042-021-11391-0

su mamografía previa para imitar el proceso clínico de contraste temporal[3]. Dicho modelo siamés de mamografía (entrenado como aprendizaje *one-shot* para detectar diferencias sutiles entre el estudio actual y el anterior) superó a modelos convencionales entrenados solo con la imagen actual, alcanzando una exactitud ~92% y un área bajo la curva (AUC) de 0.95 en la detección de cáncer. Los autores concluyen que integrar la información de la imagen previa mediante una arquitectura siamesa mejora la detección de tumores pequeños o sutiles en comparación con analizar una única imagen aislada. Estos ejemplos en mamografía y termografía ilustran que las redes siamesas son especialmente eficaces cuando: (1) existe una referencia comparativa (ya sea la otra mama, o un estudio previo del mismo paciente), y (2) los datos son escasos, pues el aprendizaje de diferencia relativa requiere menos datos absolutos que un modelo discriminativo tradicional.

Sin embargo, un punto a notar es que, si bien los modelos siameses pueden detectar la presencia de anomalías mediante un puntaje de disimilitud, no proporcionan directamente una segmentación o localización explícita de la lesión. En termografía, la salida típica de un siamés es un valor o clase indicando si el par de imágenes es "similar" (normal) o "diferente" (sugiere tumor). Para identificar la región afectada, harían falta pasos adicionales de análisis. Algunos autores han propuesto algoritmos complementarios que, una vez detectada la asimetría, tratan de inferir la ubicación de la anomalía combinando, por ejemplo, mapas térmicos o diferencias pixel a pixel entre ambas mamas. No obstante, esto no forma parte intrínseca de la arquitectura de una red siamesa estándar. En síntesis, las redes siamesas ofrecen una detección robusta de anomalías basándose en comparaciones directas (p.ej. hallando diferencias contralaterales en termogramas), pero no generan por sí solas mapas de segmentación de la lesión tumoral.

2. Uso de redes Siamesas como baseline

Las redes neuronales siamesas, que consisten en redes gemelas que comparten pesos, se han utilizado ampliamente como modelos de referencia para ciertas tareas de imágenes médicas que implican la comparación de dos imágenes relacionadas. Estas redes sobresalen en el aprendizaje de una métrica de similitud entre pares de imágenes, lo que las hace muy adecuadas para escenarios como análisis bilaterales o escaneos secuenciales.

Por ejemplo, la termografía mamaria a menudo se basa en la detección de asimetrías entre la mama izquierda y la derecha: Dey y Rajan (2024) introducen un modelo de red

[3] Dey, A., & Rajan, S. (2024). Thermography-based Breast Abnormality Detection using Siamese Network. *CMBES Proceedings*, *46*. Retrieved from https://proceedings.cmbes.ca/index.php/proceedings/article/view/1186

siamesa para conocer la similitud entre las imágenes térmicas bilaterales de mama con el fin de identificar anomalías[4]. En la mamografía, los radiólogos comparan rutinariamente las imágenes actuales y anteriores; en consecuencia, Bai et al. (2022) propusieron una CNN siamesa mejorada para detectar el cáncer de mama mediante la fusión de características de mamografías anteriores y actuales[5]. Su enfoque superó a varios modelos de referencia, incluidas las CNN de una sola imagen (ResNet, VGG) e incluso una red siamesa "vainilla", en términos de precisión y AUC [6]. Esto indica que un modelo de red siamesa bien diseñada como baseline puede ser bastante robusta, a menudo superando a los modelos más simples de una sola imagen mediante la explotación de información comparativa.

3. Transformers en la clasificación de cáncer de mama

En el contexto más general de imágenes médicas de mama (p. ej. mamografías, ecografías, histopatología), los transformers también han cobrado interés por su capacidad de auto-atención global. Un desafío conocido es que los ViT típicamente requieren grandes volúmenes de datos para entrenar desde cero; con conjuntos pequeños tienden a sobreajustar y rendir peor que las CNN[7]. En tareas de cáncer de mama, donde a menudo el número de imágenes es limitado, una estrategia exitosa ha sido usar modelos ViT pre-entrenados. Abimouloud et al. (2024) demostraron que afinando un ViT pre-entrenado en ImageNet sobre mamografías se podían obtener resultados extraordinarios: ~99.96% de exactitud en distinguir tumores benignos vs. malignos en el dataset DDSM[8]. Destacan que entrenar un ViT desde cero bajo datos escasos daba peor desempeño que redes CNN, pero con transfer learning el ViT superó dicha brecha. Este experimento ilustra cómo los transformers pueden alcanzar o incluso exceder el rendimiento de CNN en diagnóstico mamográfico, siempre que se aproveche el conocimiento aprendido en grandes datasets generales.

Otra línea de trabajo ha explorado variantes de transformers más eficientes en visión médica. Por ejemplo, los Swin-Transformers (transformers jerárquicos con ventanas desplazadas) han mostrado ser muy efectivos. Liu *et al*. aplicaron un *ensemble* de

---

[4] [4] Dey, A., & Rajan, S. (2024). Thermography-based Breast Abnormality Detection using Siamese Network. *CMBES Proceedings*, *46*. Retrieved from https://proceedings.cmbes.ca/index.php/proceedings/article/view/1186

[5] Bai, J., Jin, A., Wang, T., Yang, C., & Nabavi, S. (2022). Feature fusion Siamese network for breast cancer detection comparing current and prior mammograms. *Medical physics*, *49*(6), 3654–3669. https://doi.org/10.1002/mp.15598

[6] Ídem.

[7] Abimouloud, M.L., Bensid, K., Elleuch, M., Aiadi, O., Kherallah, M. (2024). Mammography Breast Cancer Classification Using Vision Transformers. In: Abraham, A., Bajaj, A., Hanne, T., Siarry, P. (eds) Intelligent Systems Design and Applications. ISDA 2023. Lecture Notes in Networks and Systems, vol 1046. Springer, Cham. https://doi.org/10.1007/978-3-031-64813-7_44

[8] Ídem.

modelos Swin para clasificar imágenes histopatológicas de mama en múltiples subtipos (4 subclases benignas y 4 malignas), logrando ~96.0% de exactitud en una clasificación de 8 clases, y hasta 99.6% en la clasificación binaria benigna/maligna[9]. Estos resultados en histopatología sugieren que la atención no sólo capta patrones globales, sino que puede discriminar detalles sutiles de distintas variantes tumorales. De modo similar, existen estudios en ultrasonido de mama que incorporan transformers; p. ej., Zhu *et al*. integraron Swin-Transformer a un sistema CAD en ecografía mamaria con mejoras en la detección de lesiones[10]. En general, la literatura reciente muestra que los transformers están emergiendo como modelos de alto rendimiento en imágenes médicas de mama, rivalizando con las CNN en clasificación, detección y segmentación[11].

Es importante señalar que algunos trabajos también han incorporado mecanismos de atención en arquitecturas CNN tradicionales para aprovechar sus beneficios. Por ejemplo, Alshehri y AlSaeed (2023) emplearon una CNN pre-entrenada (VGG16) complementada con módulos de *attention* para el diagnóstico en termografías, alcanzando leves mejoras: la exactitud de prueba pasó de 99.18% con VGG16 básica a 99.80% al añadir atención[12]. Si bien ambas cifras son muy altas (posiblemente debido a un dataset pequeño o balanceado), este incremento sugiere que la atención ayuda a resaltar características térmicas sutiles relevantes. En síntesis, los transformers se han validado en múltiples modalidades de cáncer de mama mostrando que, con suficiente datos o pre-entrenamiento, pueden superar a las CNN clásicas en precisión de clasificación.

4. Vision Transformers (ViT) en imágenes termográficas de mama

La termografía mamaria es una técnica de imagen infrarroja no invasiva que mide el perfil de temperatura en la superficie del seno. Se ha investigado como método complementario de screening capaz de detectar alteraciones vasculares asociadas al cáncer de mama antes de que sean visibles por mamografía convencional. Sin embargo, el análisis de termogramas es desafiante y se han aplicado técnicas de deep learning para mejorar su interpretación automática. Tradicionalmente, las redes neuronales

[9] Tummala, S., Kim, J., & Kadry, S. (2022). BreaST-Net: Multi-Class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers. *Mathematics*, *10*(21), 4109. https://doi.org/10.3390/math10214109

[10] Zhu, C., Chai, X., Xiao, Y., Liu, X., Zhang, R., Yang, Z., & Wang, Z. (2024). Swin-Net: A Swin-Transformer-Based Network Combing with Multi-Scale Features for Segmentation of Breast Tumor Ultrasound Images. *Diagnostics (Basel, Switzerland)*, *14*(3), 269. https://doi.org/10.3390/diagnostics14030269

[11] Tanimola, O., Shobayo, O., Popoola, O., & Okoyeigbo, O. (2024). Breast Cancer Classification Using Fine-Tuned SWIN Transformer Model on Mammographic Images. *Analytics*, *3*(4), 461-475. https://doi.org/10.3390/analytics3040026

[12] Alshehri, A., & AlSaeed, D. (2023). Breast Cancer Diagnosis in Thermography Using Pre-Trained VGG16 with Deep Attention Mechanisms. *Symmetry*, *15*(3), 582. https://doi.org/10.3390/sym15030582

convolucionales (CNN) han dominado estas tareas, pero recientemente los Vision Transformers (ViT) han emergido como alternativa prometedora en visión por computador médica[13].

En la literatura, existen algunos trabajos pioneros que aplican modelos tipo transformer directamente sobre imágenes térmicas de mama. Garia y Hariharan (2023) reportan uno de los primeros estudios en clasificar termogramas de mama usando *Vision Transformers*, mostrando que estos modelos de auto-atención alcanzan rendimientos *equiparables e incluso superiores* a los de CNN convencionales en la detección de lesiones malignas. En su experimento, emplearon dos variantes de ViT para clasificar imágenes termográficas en benignas vs. malignas, obteniendo altas precisiones comparables al estado del arte con CNN. Estos resultados sugieren que los transformers pueden manejar adecuadamente las características térmicas distintivas del cáncer de mama[14].

Otro enfoque relevante es el de Mahoro y Akhloufi (2022), quienes combinaron transformers y CNN en un sistema integral de detección sobre termogramas[15]. En su método, un modelo transformer de segmentación (TransUNet, basado en auto-atención) primero segmentó la región de la mama en la termografía, para luego aplicar CNN pre-entrenadas a la clasificación del tejido segmentado. Probaron cuatro arquitecturas (EfficientNet-B7, ResNet-50, VGG-16 y DenseNet-201) para clasificar cada termograma segmentado en *sano*, *enfermo* o *desconocido*, utilizando 3989 imágenes de la base de datos DMR-IR. La mejor precisión lograda es del 97,26%, la sensibilidad del 97,26% y la especificidad de cada clase sano, enfermo, desconocido es del 100%, 96,94% y 99,72% respectivamente con el modelo ResNet-50[16].

Aunque en este trabajo el transformer (TransUNet) se usó principalmente para segmentación y las CNN para clasificación, demuestra la viabilidad de integrar transformers en la cadena de procesamiento de termografías mamarias. El preprocesamiento mediante segmentación parece útil, dado que enfoca el análisis en la zona mamaria relevante, eliminando fondo y otras regiones corporales.

En resumen, si bien los estudios sobre Transformers aplicados directamente a imágenes termográficas de mama aún son escasos, los resultados iniciales son prometedores. Los ViT han logrado desempeños cercanos o superiores a CNN en clasificación binaria de

---

[13] Garia, L.S., Hariharan, M. (2023). Vision Transformers for Breast Cancer Classification from Thermal Images. In: Muthusamy, H., Botzheim, J., Nayak, R. (eds) Robotics, Control and Computer Vision. Lecture Notes in Electrical Engineering, vol 1009. Springer, Singapore. https://doi.org/10.1007/978-981-99-0236-1_13

[14] Ídem.

[15] Mahoro E., Akhloufi M.A. Breast cancer classification on thermograms using deep CNN and transformers. Quant. Infrared Thermogr. J. 2022:1–20. doi: 10.1080/17686733.2022.2129135.

[16] Ídem.

termogramas[17], y esquemas híbridos Transformer+CNN han mostrado altas precisiones en detección de cáncer térmico. Esto sienta un precedente positivo para explorar transformers en esta modalidad de imagen.

5. Detection Transformer en imágenes médicas

El modelo Detection Transformer (DETR), introducido por Carion et al. (2020)[18], representa un enfoque end-to-end para la detección de objetos mediante mecanismos de autoatención, eliminando la necesidad de etapas intermedias como la generación de propuestas. En el dominio médico, se han adaptado variantes de DETR para la detección de lesiones en diferentes modalidades. Por ejemplo, algunos estudios han explorado su uso en la detección de pólipos en imágenes endoscópicas y en la segmentación de lesiones en tomografías, demostrando que el enfoque de atención global es capaz de capturar relaciones espaciales complejas incluso en imágenes con características sutiles y de bajo contraste.

Estos estudios sugieren que, con una adecuada adaptación al preprocesamiento de las termografías y a la naturaleza monocromática de las imágenes térmicas, DETR podría usarse para detectar áreas sospechosas de malignidad en imágenes sintéticas generadas por GANs, siempre que se realice un cuidadoso ajuste (fine-tuning) y se cuente con suficientes datos de entrenamiento, reales y sintéticos.

6. SegFormer en tareas de segmentación

Por otro lado, SegFormer, propuesto por Xie et al. (2021)[19], es un modelo de segmentación semántica que combina la eficiencia de los transformers con la capacidad de capturar contextos multiescalares sin recurrir a estructuras jerárquicas complejas. Su arquitectura, que no depende de convoluciones en etapas finales, ha demostrado ser eficaz en la segmentación de estructuras complejas en imágenes médicas, como en estudios de segmentación de lesiones en tomografías y resonancias magnéticas.

Aunque la aplicación específica de SegFormer a imágenes termográficas de mama aún está en exploración, los resultados obtenidos en otros campos sugieren que el modelo puede adaptarse para segmentar regiones relevantes (por ejemplo, áreas con anomalías

---

[17] Garia, L.S., Hariharan, M. (2023). Vision Transformers for Breast Cancer Classification from Thermal Images. In: Muthusamy, H., Botzheim, J., Nayak, R. (eds) Robotics, Control and Computer Vision. Lecture Notes in Electrical Engineering, vol 1009. Springer, Singapore. https://doi.org/10.1007/978-981-99-0236-1_13

[18] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). *End-to-End object detection with transformers*. arXiv. https://arxiv.org/abs/2005.12872

[19] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). *SegFormer: Simple and efficient design for semantic segmentation with transformers*. arXiv. https://arxiv.org/abs/2105.15203

térmicas asociadas a tumores) en imágenes sintéticas. Esto permitiría, por ejemplo, delimitar de forma precisa la zona de interés antes de proceder a una clasificación final con otro modelo o incluso integrarlo en un sistema de detección end-to-end.

7. Comparación de Vision Transformers (ViT), Detection Transformers (DETR) y SegFormer

7.1 Tabla comparativa de ViT, DETR y SegFormer

Características	Vision Transformers (ViT)	Detection Transformers (DETR)	SegFormer
Función Principal	Clasificación global (benigno vs. maligno)	Detección de objetos con predicción de bounding boxes	Segmentación semántica (delimitación de regiones de interés)
Arquitectura	Encoder transformer aplicado sobre patches	CNN backbone + encoder-decoder transformer (atención global)	Transformer jerárquico + decodificador MLP ligero
Preprocesamiento Requerido	Redimensionamiento, normalización, división en patches, embeddings	Normalización, redimensionamiento, extracción de características (CNN)	Redimensionamiento, normalización, posible alineación y estandarización de resolución
Ventajas	Captura relaciones globales; ideal para clasificación con transfer learning	Arquitectura end-to-end; detección sin necesidad de propuestas o NMS	Segmentación precisa y eficiente; manejo de información multiescala
Limitaciones	Requiere gran cantidad de datos o pre-entrenamiento; no realiza detección o segmentación directa	Mayor requerimiento de datos y tiempo de entrenamiento; precisión en bounding boxes puede verse afectada	Dependencia de anotaciones precisas a nivel de píxel; sensibilidad a artefactos en imágenes sintéticas
Aplicabilidad al Caso Propuesto	Ideal para clasificación final si se dispone de ROI presegmentadas	Adecuado para localizar lesiones con bounding boxes	Muy útil para delimitar áreas tumorales y extraer contornos para análisis posterior

7.2 Conclusiones sobre uso de Tranformers en imágenes termográficas

Para la detección, segmentación y clasificación de lesiones tumorales mamarias en imágenes termográficas sintéticas mediante GAN, la elección óptima dependerá del enfoque deseado:

- **Segmentación y extracción de regiones de interés:** SegFormer es especialmente adecuado para delimitar con precisión las lesiones, facilitando posteriores análisis morfológicos o de clasificación.
- **Detección end-to-end:** DETR ofrece una solución robusta para localizar y clasificar las lesiones mediante predicción directa de bounding boxes, siendo ideal si se requiere una integración de localización y clasificación en un solo modelo.
- **Clasificación global:** ViT puede ser efectivo para la clasificación final de las lesiones (benigno vs. maligno) si se dispone de un preprocesamiento que asegure que las imágenes de entrada se centren en las áreas de interés, especialmente aprovechando transfer learning.

Cada uno de estos modelos presenta desafíos específicos en términos de preprocesamiento y requisitos de datos, pero la integración de datos sintéticos generados por GAN puede mitigar la limitación de datos reales y mejorar la generalización. Un pipeline combinado que aproveche las fortalezas de cada arquitectura – por ejemplo, usando SegFormer para segmentación y ViT o DETR para clasificación/detección – podría resultar en un sistema robusto y de alto rendimiento para el diagnóstico asistido en termografía mamaria.

8. Viabilidad técnica de arquitecturas Transformer con GANs

Considerando lo anterior, es técnicamente factible aplicar arquitecturas transformer para clasificar lesiones mamarias benignas vs malignas en termografías, apoyándose en imágenes sintéticas generadas por GAN. Los estudios revisados brindan fundamentos optimistas: (1) Los transformers (ViT, Swin, TransUNet) ya han sido empleados con éxito tanto en termografías reales[20,21] como en otras imágenes de mama[22], logrando desempeños de primer nivel. (2) La generación de termogramas sintéticos es plausible mediante diversas GAN (ya sea para aumentar el volumen de datos o para equilibrar clases) y experiencias en dominios afines demuestran que dichos datos pueden mejorar sustancialmente la precisión de clasificación[23]. Integrar ambas tecnologías podría, por tanto, potenciar un sistema CAD de detección de cáncer de mama.

No obstante, hay consideraciones importantes para asegurar la viabilidad y mejorar el rendimiento:

8.1 **Calidad y realismo de los datos sintéticos**

La efectividad de entrenar transformers con imágenes sintéticas dependerá de qué tan realistas sean éstas. Las GAN propuestas ofrecen diferentes ventajas: las WGAN (Wasserstein GAN) y SNGAN (GAN con normalización espectral) tienden a una convergencia más estable y evitan artefactos de entrenamiento, generando distribuciones globalmente realistas. Las cGAN permiten incluir la *etiqueta* (benigno/maligno) como condicionante, asegurando que las imágenes simuladas reflejen características propias de cada clase. Por su parte, las CycleGAN aprenden transformaciones no emparejadas entre dominios, útiles para conservar la estructura anatómica al intercambiar sólo las propiedades térmicas patológicas. Por ejemplo, un CycleGAN podría aprender a tomar un

[20] Mahoro, E., & Akhloufi, M. A. (2022). Applying Deep Learning for Breast Cancer Detection in Radiology. *Current oncology (Toronto, Ont.)*, *29*(11), 8767–8793. https://doi.org/10.3390/curroncol29110690

[21] Garia, L.S., Hariharan, M. (2023). Vision Transformers for Breast Cancer Classification from Thermal Images. In: Muthusamy, H., Botzheim, J., Nayak, R. (eds) Robotics, Control and Computer Vision. Lecture Notes in Electrical Engineering, vol 1009. Springer, Singapore. https://doi.org/10.1007/978-981-99-0236-1_13

[22] Abimouloud, M.L., Bensid, K., Elleuch, M., Aiadi, O., Kherallah, M. (2024). Mammography Breast Cancer Classification Using Vision Transformers. In: Abraham, A., Bajaj, A., Hanne, T., Siarry, P. (eds) Intelligent Systems Design and Applications. ISDA 2023. Lecture Notes in Networks and Systems, vol 1046. Springer, Cham. https://doi.org/10.1007/978-3-031-64813-7_44

[23] Yurtsever, M. M. E., Atay, Y., Arslan, B., & Sagiroglu, S. (2024). Development of brain tumor radiogenomic classification using GAN-based augmentation of MRI slices in the newly released gazi brains dataset. *BMC medical informatics and decision making*, *24*(1), 285. https://doi.org/10.1186/s12911-024-02699-6

pecho sano e insertar un patrón térmico maligno plausible[24], creando un caso artificial pero coherente anatómicamente. La literatura sugiere que CycleGAN produce imágenes de mayor fidelidad visual en tareas médicas específicas[25], por lo que podría ser valiosa para generar termogramas tumores con pocos artefactos. En cualquier caso, será crucial validar visualmente que las imágenes sintéticas imiten correctamente las distribuciones de temperatura de tumores reales. Imágenes sintéticas de baja calidad podrían confundir más que ayudar o introducir sesgos indeseados en el transformer.

8.2 **Volumen de datos y pre-entrenamiento**

Aun con datos sintéticos, el conjunto efectivo para entrenamiento debe ser suficientemente amplio y variado. Los transformers tienen muchos parámetros y menos sesgo inductivo que las CNN, por lo que son más propensos a sobreajustar con pocos ejemplos[26]. Incorporar miles de termogramas generados por GAN podría mitigar este problema al exponer al modelo a más variaciones. Además, adoptar un enfoque de pre-entrenamiento sería recomendable. Por ejemplo, utilizar un ViT inicialmente entrenado en un gran conjunto de imágenes naturales o médicas (aunque no sean térmicas) provee representaciones generales que luego se ajustan a las termografías. Alternativamente, se podría pre-entrenar el transformer directamente sobre las imágenes sintéticas (fácilmente generables en gran cantidad) antes de fine-tuning en las imágenes reales etiquetadas. Esta metodología de usar primero los datos sintéticos y luego los podría aprovechar lo mejor de ambos mundos: el volumen de datos sintéticos y la fidelidad de los datos reales, reduciendo la dependencia exclusiva en cada uno.

8.3 **Estrategia de entrenamiento con datos sintéticos**

Se debe decidir cómo mezclar los datos reales y sintéticos. Una práctica común es la aumentación por sobresampling, donde cada minibatch de entrenamiento contiene tanto imágenes reales como algunas sintéticas nuevas (quizá en proporción

---

[24] Becker, A. S., Jendele, L., Skopek, O., Berger, N., Ghafoor, S., Marcon, M., & Konukoglu, E. (2019). Injecting and removing suspicious features in breast imaging with CycleGAN: A pilot study of automated adversarial attacks using neural networks on small images. *European journal of radiology*, *120*, 108649. https://doi.org/10.1016/j.ejrad.2019.108649

[25] Jiménez-Gaona, Y., Carrión-Figueroa, D., Lakshminarayanan, V., Rodríguez-Álvarez, M.J. (2024). Gan-based data augmentation to improve breast ultrasound and mammography mass classification, Biomedical Signal Processing and Control, Volume 94, 2024, 106255, ISSN 1746-8094. https://doi.org/10.1016/j.bspc.2024.106255

[26] Abimouloud, M.L., Bensid, K., Elleuch, M., Aiadi, O., Kherallah, M. (2024). Mammography Breast Cancer Classification Using Vision Transformers. In: Abraham, A., Bajaj, A., Hanne, T., Siarry, P. (eds) Intelligent Systems Design and Applications. ISDA 2023. Lecture Notes in Networks and Systems, vol 1046. Springer, Cham. https://doi.org/10.1007/978-3-031-64813-7_44

equilibrada de clases). Es importante evitar que el modelo aprenda a distinguir lo sintético de lo real en vez de las patologías; para ello, la calidad de las GAN es clave, pero también se pueden introducir las imágenes generadas gradualmente. Otra táctica es usar los sintéticos principalmente para pre-entrenar o para expandir la clase minoritaria (p. ej. generar muchos malignos si escasean), mientras que la evaluación final se hace solo en datos reales. La combinación óptima puede requerir experimentación. Estudios previos han mostrado que la ganancia de rendimiento al usar datos sintéticos es significativa cuando la data real es limitada[27], pero tiende a saturar si las GAN comienzan a producir imágenes redundantes o de menor diversidad que las reales.

8.4 **Características de las termografías**

Las imágenes térmicas de mama tienen particularidades a considerar. Su resolución puede ser relativamente baja (comparada con foto óptica o radiografía) y suelen ser imágenes monocromáticas (mapas de calor). Un ViT estándar operando con patches de, por ejemplo, 16×16 píxeles podría funcionar bien dado que las estructuras térmicas relevantes (zonas de hipercalor) suelen ser regiones más difusas pero abarcativas. La auto-atención podría identificar correlaciones simétricas entre ambas mamas o patrones vasculares globales que distingan un tumor. De hecho, una ventaja del transformer es que captura relaciones de largo alcance: en termografía, una señal importante de malignidad es la asimetría térmica entre el seno izquierdo y derecho o entre zonas del mismo seno. Un modelo de atención podría comparar distintas partes de la imagen entre sí de forma más directa que una CNN pura. Claro está, si la adquisición incluye ambos senos en la misma imagen, el modelo podría aprender a comparar; si no, podría alimentarse al modelo una imagen compuesta de ambas vistas. En cualquier caso, ya que los transformers tratan la imagen como una secuencia de tokens, pueden integrar información distribuida en el cuadro termográfico entero.

8.5 **Segmentación**

Como mostró Mahoro et al., segmentar la región de la mama puede aumentar la precisión[28]. Un paso de segmentación automática (posiblemente también con un modelo basado en transformer, como TransUNet) podría incorporarse al *pipeline*:

---

[27] Alzahem, A., Boulila, W., Koubaa, A. *et al.* Improving satellite image classification accuracy using GAN-based data augmentation and vision transformers. *Earth Sci Inform* **16**, 4169–4186 (2023). https://doi.org/10.1007/s12145-023-01153-x

[28] Mahoro, E., & Akhloufi, M. A. (2022). Applying Deep Learning for Breast Cancer Detection in Radiology. *Current oncology (Toronto, Ont.)*, *29*(11), 8767–8793. https://doi.org/10.3390/curroncol29110690

primero aislar la mama del torso y fondo, luego clasificar la región aislada. Esto removería posibles "falsos calientes" ajenos (por ejemplo, la zona del corazón o axilas) que pudieran distraer al clasificador. Alternativamente, el transformer mismo podría entrenarse para ignorar el fondo si se le proporcionan suficientes ejemplos; su atención debería centrarse en las áreas con tejido mamario. No obstante, la segmentación previa es una técnica recomendada para termografías según varios estudios, ya que normaliza la entrada (cada imagen se centra solo en la glándula mamaria).

8.6 **Capacidad de generalización**

Un riesgo de usar datos sintéticos es que el modelo aprenda *artefactos* de generación y no generalice bien a imágenes de pacientes reales inéditos. Por ello, se debe evaluar el sistema en un conjunto de pruebas independiente de imágenes reales sin ninguna mezcla sintética. Idealmente, los modelos resultantes deberían ser comparados contra métodos previos. Dado que ya existen enfoques CNN que alcanzan >95% de acierto en termografía[29], la meta sería comprobar si un transformer + datos sintéticos puede superar consistentemente ese rendimiento (por ejemplo, acercarse más al 98-99% con buen equilibrio de sensibilidad/especificidad). En la literatura, cuando se han usado GANs para aumentación, se ha logrado reducir *overfitting* y mejorado la robustez del clasificador a variaciones de las lesiones[30]. Por tanto, es de esperar que un transformer entrenado con abundantes termogramas sintéticos vea reforzada su capacidad de generalizar a distintas presentaciones de tumores (tamaños, ubicaciones, intensidades térmicas, etc.).

En conclusión, el enfoque de combinar transformers con imágenes termográficas sintéticas mediante GAN es viable y respaldado por tendencias actuales. Los transformers aportan un marco potente para la clasificación de imágenes médicas complejas, habiendo demostrado eficacia tanto en termogramas[31] como en otras

---

[29] Mahoro, E., & Akhloufi, M. A. (2022). Applying Deep Learning for Breast Cancer Detection in Radiology. *Current oncology (Toronto, Ont.)*, *29*(11), 8767–8793. https://doi.org/10.3390/curroncol29110690

[30] Yurtsever, M. M. E., Atay, Y., Arslan, B., & Sagiroglu, S. (2024). Development of brain tumor radiogenomic classification using GAN-based augmentation of MRI slices in the newly released gazi brains dataset. *BMC medical informatics and decision making*, *24*(1), 285. https://doi.org/10.1186/s12911-024-02699-6

[31] Garia, L.S., Hariharan, M. (2023). Vision Transformers for Breast Cancer Classification from Thermal Images. In: Muthusamy, H., Botzheim, J., Nayak, R. (eds) Robotics, Control and Computer Vision. Lecture

modalidades de cáncer de mama[32]. Por su parte, las GANs ofrecen una solución al problema de datos escasos, permitiendo generar ejemplos realistas de lesiones para enriquecer el entrenamiento[33]. La convergencia de ambas tecnologías se alinea con los avances recientes en aprendizaje profundo médico.

Basándonos en la revisión de la literatura, es razonable esperar que un ViT (o Swin) entrenado sobre una combinación de termogramas reales y sintéticos logre al menos igualar, si no mejorar, el estado del arte en clasificación binaria de cáncer de mama por termografía. Eso sí, el éxito dependerá de ejecutar cuidadosamente cada componente: una correcta generación y utilización de los datos sintéticos, un adecuado esquema de entrenamiento (posiblemente con pre-entrenamiento) y la consideración de las características propias de la termografía. Con rigor técnico en esas áreas, mejorar el rendimiento de un sistema asistido de diagnóstico de cáncer de mama mediante transformers + GAN es altamente factible, apoyado en fundamentos científicos claros y en resultados preliminares alentadores en la literatura.

Notes in Electrical Engineering, vol 1009. Springer, Singapore. https://doi.org/10.1007/978-981-99-0236-1_13

[32] Abimouloud, M.L., Bensid, K., Elleuch, M., Aiadi, O., Kherallah, M. (2024). Mammography Breast Cancer Classification Using Vision Transformers. In: Abraham, A., Bajaj, A., Hanne, T., Siarry, P. (eds) Intelligent Systems Design and Applications. ISDA 2023. Lecture Notes in Networks and Systems, vol 1046. Springer, Cham. https://doi.org/10.1007/978-3-031-64813-7_44

[33] Yurtsever, M. M. E., Atay, Y., Arslan, B., & Sagiroglu, S. (2024). Development of brain tumor radiogenomic classification using GAN-based augmentation of MRI slices in the newly released gazi brains dataset. *BMC medical informatics and decision making*, *24*(1), 285. https://doi.org/10.1186/s12911-024-02699-6

9. Datos sintéticos generados por GANs para uso con Transformers en imágenes médicas

Un obstáculo común en el entrenamiento de modelos profundos en imágenes médicas es la escasez de datos y el desequilibrio entre clases (p. ej., relativamente pocos casos malignos confirmados versus muchos casos benignos). Aquí es donde las Redes Generativas Antagónicas (GANs) aportan una herramienta valiosa: la generación de *datos sintéticos* que aumenten y diversifiquen el conjunto de entrenamiento. La idea de combinar GANs para ampliar datos junto con modelos de clasificación (incluyendo transformers) ha ganado terreno en años recientes[34].

Existen trabajos generales que demuestran la eficacia de esta sinergia. Por ejemplo, Alzahem et al. (2023) presentaron un esquema de aumentación de datos con GAN asistido por Vision Transformer en imágenes satelitales[35]. En su enfoque, primero se entrena una GAN para generar imágenes sintéticas con la misma distribución estadística que las reales, y luego un ViT aprovecha este conjunto expandido para la clasificación. Los resultados mostraron una mejora drástica de rendimiento: la exactitud pasó de 76.9% (usando sólo aumentación tradicional) a 98.7% empleando la combinación GAN+ViT[36]. Aunque esto fue en teledetección, es una prueba de concepto contundente de que los datos sintéticos pueden aumentar significativamente la capacidad predictiva de un transformer en clasificación de imágenes.

En el dominio médico, se han visto beneficios similares. Un estudio sobre tumores cerebrales (MRI) utilizó StyleGAN2-ADA para generar cortes sintéticos de resonancia, añadiéndolos al entrenamiento de un clasificador de tumores[37]. Como resultado, los modelos (EfficientNet) mejoraron notablemente su precisión global en la detección de tumores cerebrales en múltiples datasets, alcanzando hasta 99.36% de exactitud en algunos casos[38]. Los autores enfatizan *la potencia de los GANs* para aumentar conjuntos

---

[34] Yurtsever, M. M. E., Atay, Y., Arslan, B., & Sagiroglu, S. (2024). Development of brain tumor radiogenomic classification using GAN-based augmentation of MRI slices in the newly released gazi brains dataset. *BMC medical informatics and decision making*, *24*(1), 285. https://doi.org/10.1186/s12911-024-02699-6

[35] Alzahem, A., Boulila, W., Koubaa, A. *et al.* Improving satellite image classification accuracy using GAN-based data augmentation and vision transformers. *Earth Sci Inform* **16**, 4169–4186 (2023). https://doi.org/10.1007/s12145-023-01153-x

[36] Ídem.

[37] Yurtsever, M. M. E., Atay, Y., Arslan, B., & Sagiroglu, S. (2024). Development of brain tumor radiogenomic classification using GAN-based augmentation of MRI slices in the newly released gazi brains dataset. *BMC medical informatics and decision making*, *24*(1), 285. https://doi.org/10.1186/s12911-024-02699-6

[38] Ídem.

de datos médicos, particularmente en clasificación de tumores, mostrando incrementos notables de exactitud mediante la integración de datos sintéticos[39].

Esto sugiere que, incluso en aplicaciones de salud, las imágenes generadas por GAN (si son lo suficientemente realistas) pueden robustecer el entrenamiento de modelos y elevar su desempeño en detección de cáncer.

En el caso específico de cáncer de mama, también hay exploraciones de GANs para datos sintéticos. Jiménez-Gaona *et al*. (2024) estudiaron la aumentación en ecografías y mamografías de tumores mamarios usando múltiples arquitecturas GAN (DCGAN, CycleGAN, WGAN-GP, etc.) para equilibrar clases. Hallaron que no todas las GAN producen resultados igual de útiles: por ejemplo, CycleGAN generó imágenes con mayor similitud visual a las reales, mientras que un WGAN-GP produjo imágenes de menor calidad con artefactos apreciables[40]. Esto indica que la selección de la arquitectura GAN es importante; modelos de transferencia de estilo como CycleGAN pueden preservar mejor las características anatómicas relevantes (p. ej., textura de lesiones) en comparación con GAN estándar con restricción de gradiente. Además, se ha propuesto utilizar GANs condicionales (cGAN) para dirigir la generación de imágenes con una etiqueta dada; por ejemplo, crear sintéticamente termogramas *malignos* o *benignos* según se necesite, aumentando específicamente la clase minoritaria. Un cGAN entrenado con etiquetas podría así proporcionar ejemplos adicionales de lesiones malignas infrecuentes, ayudando a equilibrar el dataset para el transformer. Incluso es posible entrenar *dos* GAN separadas (o una cGAN bipartita) para cada clase de termograma, logrando muestras sintéticas etiquetadas. De este modo, las imágenes sintéticas pueden integrarse directamente en el entrenamiento supervisado de un ViT binario benigno/maligno.

10. Arquitectura SegFormer + Vision Transformers en imágenes médicas

10.1 Estado del arte

Los modelos transformer han demostrado gran potencial en imágenes médicas tanto para segmentación semántica como para clasificación. En particular, se han explorado pipelines que integran un modelo de segmentación tipo SegFormer (encoder transformer jerárquico para segmentación) junto con modelos de Vision Transformer (ViT) para clasificar las regiones de interés segmentadas. Diversos estudios revisados por pares

[39] Ídem.
[40] Jiménez-Gaona, Y., Carrión-Figueroa, D., Lakshminarayanan, V., Rodríguez-Álvarez, M.J. (2024). Gan-based data augmentation to improve breast ultrasound and mammography mass classification, Biomedical Signal Processing and Control, Volume 94, 2024, 106255, ISSN 1746-8094. https://doi.org/10.1016/j.bspc.2024.106255.

han aplicado esta combinación en distintos dominios médicos (cerebro, mama, colon, patología digital, etc.), aprovechando la segmentación precisa de lesiones/anatomía y la posterior clasificación basada en atención global. A continuación, se resumen ejemplos representativos respaldados por la literatura:

- **Tumores cerebrales (MRI):** Rahman *et al*. (2024) investigaron un pipeline con SegFormer y otros transformers para análisis de tumores cerebrales. Usaron SegFormer afinado para segmentar tumores en MRI, y un modelo transformer ConvNeXt V2 para clasificarlos, comparando su desempeño con Vision Transformers clásicos. El esquema logró segmentaciones precisas (Dice ≈90%) y clasificación casi perfecta (99.6% de exactitud), superando a ViT puro en dicha tarea[41]. Este estudio demuestra la viabilidad de combinar un transformer de segmentación con otro de clasificación en neuro-oncología.

- **Ultrasonido de mama (BUS):** Tagnamas *et al*. (2024) propusieron un marco *multi-task* para segmentar y luego clasificar tumores de mama en ecografías. Primero segmentan la lesión utilizando un encoder dual (un CNN EfficientNetV2 + un encoder ViT adaptado) fusionados con atención de canales, obteniendo el mapa segmentado[42]. Luego, clasifican la región tumoral segmentada (benigno vs maligno) mediante un clasificador MLP-Mixer ligero (evaluaron también un ViT para este paso). Su enfoque superó trabajos previos, logrando 83.42% en segmentación y 86% exactitud en clasificación, evidenciando las ventajas de integrar Transformers en ambos sub-procesos.

- **Pólipos en colonoscopia:** Li *et al*. (2024) desarrollaron una red *dual-branch* de segmentación + clasificación simultáneas para pólipos colorectales. La arquitectura consta de una *rama transformer* y una *rama CNN*, explotando la atención global del transformer junto con el detalle local del CNN[43]. Un módulo de interacción de características (FIM) integra las representaciones de ambas ramas, capturando completamente el contexto global y detalles locales relevantes. Adicionalmente, incorporan un módulo de realce de bordes por atención invertida (RABE) para refinar contornos, mejorando la detección de los límites del pólipo. Probada en cinco conjuntos de datos públicos, esta solución

---

[41] M. A. Rahman, A. Joy, A. T. Abir and T. Shimamura, "Unleashing the Power of Open-Source Transformers in Medical Imaging: Insights from a Brain" International Journal of Advanced Computer Science and Applications(IJACSA), 15(7), 2024. http://dx.doi.org/10.14569/IJACSA.2024.01507126

[42] Tagnamas, J., Ramadan, H., Yahyaouy, A. *et al.* Multi-task approach based on combined CNN-transformer for efficient segmentation and classification of breast tumors in ultrasound images. *Vis. Comput. Ind. Biomed. Art* **7**, 2 (2024). https://doi.org/10.1186/s42492-024-00155-w

[43] Chenqian Li, Jun Liu, Jinshan Tang. Simultaneous segmentation and classification of colon cancer polyp images using a dual branch multi-task learning network[J]. Mathematical Biosciences and Engineering, 2024, 21(2): 2024-2049. doi: 10.3934/mbe.2024090

multi-tarea logró resultados superiores a los *state of the art* (métodos estado del arte) en segmentación y clasificación de pólipos.

- **Histopatología digital (células):** Hörst *et al*. (2024) introdujeron CellViT, una arquitectura basada en Vision Transformers para segmentación *instance* de núcleos celulares con clasificación del tipo celular simultánea. Entrenado en el exigente dataset PanNuke (≈200k núcleos anotados en 5 clases celulares distintas), CellViT aprovecha un encoder ViT pre-entrenado masivamente en imágenes histológicas y el modelo de segmentación generalista SAM para mejorar su desempeño[44]. El resultado fue **estado del arte** en detección y segmentación de núcleos (Panoptic Quality 0.50; F1 detección 0.83), mostrando la eficacia de los Transformers para segmentar regiones celulares y clasificar su categoría histológica en un solo sistema. Este trabajo, publicado en *Medical Image Analysis*, destaca el poder de combinar segmentación por transformers con clasificación de instancias en patología computacional.

10.2 Vision Transformers y SegFormer en análisis de imágenes médicas

Frente al enfoque de redes siamesas (basado usualmente en CNN dobles), han emergido en años recientes arquitecturas basadas en Transformers de Visión para tareas de análisis de imágenes, incluidas las médicas. Un Vision Transformer (ViT) aplica el mecanismo de *self-attention* sobre parches de la imagen, capturando dependencias globales sin recurrir a convoluciones tradicionales[45]. Esta arquitectura, introducida originalmente para clasificación general de imágenes, ha demostrado ser un potente sustituto o complemento de las CNN, especialmente cuando se dispone de grandes volúmenes de datos o modelos pre-entrenados. En el ámbito médico, ya se han reportado aplicaciones exitosas de ViT. Por ejemplo, Garia *et al*. (2023) exploraron por primera vez el uso de Vision Transformers para clasificación de imágenes térmicas de mama, encontrando que modelos ViT alcanzan una eficacia similar o superior a la de CNNs en dicha tarea. Concretamente, evaluaron dos variantes de ViT pre-entrenados y obtuvieron resultados competitivos en términos de exactitud y área bajo la curva, lo que sugiere que el *self-attention* puede capturar patrones térmicos relevantes de lesiones de mama de manera tan efectiva como las redes convolucionales clásicas. Este hallazgo es importante porque las CNN ya habían logrado buenos desempeños en termogramas

[44] Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., & Kleesiek, J. (2024). CellViT: Vision Transformers for precise cell segmentation and classification. *Medical image analysis*, *94*, 103143. https://doi.org/10.1016/j.media.2024.103143

[45] Garia, L., & Hariharan, M. (2023, May). *Vision transformers for breast cancer classification from thermal images* (pp. 177–185). https://doi.org/10.1007/978-981-99-0236-1_13

mamarios (típicamente >90% de acierto con suficientes datos), y ahora los Transformers aparecen como una alternativa viable incluso en este dominio especializado.

Para tareas de segmentación de lesiones en imágenes médicas, se han desarrollado arquitecturas basadas en Transformers como SegFormer. SegFormer (Xie *et al.*, 2021[46]) es un modelo de segmentación semántica que utiliza un encoder de Transformers eficiente y jerárquico (sin positional encodings pesados) acoplado a un decodificador ligero totalmente convolucional. En términos simples, SegFormer aprovecha la capacidad de los Transformers para capturar contexto global a múltiples escalas y produce máscaras de segmentación de alta calidad sin el complejo *upsampling* de modelos tradicionales[47]. En la literatura médica, SegFormer ha empezado a evaluarse como posible reemplazo de la conocida arquitectura U-Net (base de muchos segmentadores médicos desde 2015). Por ejemplo, un estudio comparativo (Sourget *et al.*, MIUA 2023) probó SegFormer vs. U-Net en tres tareas de segmentación médica distintas (ecografías cardíacas, pólipos en endoscopía, e instrumentos en colonoscopía) . Los resultados mostraron que SegFormer puede ser un verdadero competidor de U-Net, logrando desempeños similares en métricas de segmentación (Dice, IoU) e incluso con ventajas en eficiencia de entrenamiento. Esto indica que las arquitecturas Transformer no solo sirven para clasificar imágenes completas, sino también para segmentar estructuras anatómicas o regiones patológicas con precisión comparable al estado del arte convencional.

En el contexto de imágenes térmicas de mama, la aplicación de segmentadores tipo SegFormer aún es incipiente, pero es razonable suponer que podrían mejorar la identificación exacta de las regiones calientes asociadas a tumores. Hasta ahora, la segmentación de tumores en termogramas se ha abordado con técnicas como clustering o contornos activos en estudios previos, y más recientemente con CNN tipo encoder-decoder (U-Net)[48]. Por ejemplo, Kakileti *et al.* (2019)[49] exploraron varios modelos de segmentación en termogramas infrarrojos de mama, demostrando que un enfoque de encoder-decoder CNN superaba a clasificadores por parches en detección

---

[46] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. Wortman Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 12077–12090). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf

[47] Sourget, T., Hasany, S.N., Mériaudeau, F., Petitjean, C. (2024). Can SegFormer be a True Competitor to U-Net for Medical Image Segmentation?. In: Waiter, G., Lambrou, T., Leontidis, G., Oren, N., Morris, T., Gordon, S. (eds) Medical Image Understanding and Analysis. MIUA 2023. Lecture Notes in Computer Science, vol 14122. Springer, Cham. https://doi.org/10.1007/978-3-031-48593-0_8

[48] Garia, L., & Hariharan, M. (2023, May). *Vision transformers for breast cancer classification from thermal images* (pp. 177–185). https://doi.org/10.1007/978-981-99-0236-1_13

[49] Kakileti, S. T., Dalmia, A., & Manjunath, G. (2019, June). Exploring deep learning networks for tumour segmentation in infrared images. *Quantitative InfraRed Thermography Journal, 17*, 1–16. https://doi.org/10.1080/17686733.2019.1619355

de “hotspots” tumorales, incluso cuando el conjunto de datos era pequeño. Dado que SegFormer pertenece a esta familia de encoder-decoder (pero con Transformer en el encoder), cabe esperar al menos un desempeño equivalente o mejor, gracias a su capacidad de capturar relaciones de largo alcance en la imagen. De hecho, SegFormer y otros transformadores segmentadores ofrecen una ventaja importante: unifican la percepción global y detalle local, lo que podría ayudar a distinguir un tumor verdadero de artefactos termales dispersos en la imagen, algo crucial para reducir *falsos positivos*.

Un aspecto para considerar es que los Transformers como ViT típicamente tienen muchos parámetros y menos sesgos inductivos que las CNN, por lo que tradicionalmente requieren datasets más grandes para entrenar desde cero de forma efectiva. Sin embargo, en la práctica médica se suele aprovechar aprendizaje transferido (pre-entrenamiento en conjuntos masivos como ImageNet) seguido de *fine-tuning*, lo que mitiga la necesidad de miles de imágenes médicas de entrenamiento. El estudio de Garia *et al*. probablemente utilizó modelos ViT pre-entrenados y ajustados a los termogramas, logrando buen rendimiento sin un volumen enorme de imágenes propias. Asimismo, SegFormer puede ser inicializado con pesos entrenados en segmentación general y luego refinado en datos médicos específicos, como sugieren los experimentos de Sourget *et al*. (donde SegFormer funcionó bien con los datos de tamaño moderado de cada desafío)[50]. En resumen, los enfoques basados en ViT/SegFormer requieren (y se benefician de) más datos o pre-entrenamiento que una CNN clásica, pero cuando esa condición se satisface, ofrecen capacidades de modelado superiores en cuanto a capturar patrones complejos globales y proporcionar salidas de segmentación detalladas.

11. Comparación de ViT + SegFormer vs. Redes Siamesas

A continuación, se comparan las características de utilizar un pipeline moderno con Vision Transformer (para clasificación) y SegFormer (para segmentación) frente al enfoque tradicional de red siamés, en el contexto de detección y análisis de tumores mamarios en termografías:

- Arquitectura: La solución ViT + SegFormer procesa cada imagen de forma individual con una red de *self-attention* (ViT) para clasificarla, y con un decoder de segmentación (SegFormer u otro) para delinear la lesión. Esto significa que modela directamente la relación entre píxeles de una misma imagen,

[50] Sourget, T., Hasany, S.N., Mériaudeau, F., Petitjean, C. (2024). Can SegFormer be a True Competitor to U-Net for Medical Image Segmentation?. In: Waiter, G., Lambrou, T., Leontidis, G., Oren, N., Morris, T., Gordon, S. (eds) Medical Image Understanding and Analysis. MIUA 2023. Lecture Notes in Computer Science, vol 14122. Springer, Cham. https://doi.org/10.1007/978-3-031-48593-0_8

aprendiendo representaciones globales de la distribución térmica[51]. En cambio, una red siamesa toma pares de imágenes (seno izquierdo y derecho) y las pasa por dos ramas CNN idénticas para obtener vectores de características que luego compara[52]. La arquitectura siamesa, por diseño, requiere dos entradas correlacionadas (p.ej., los dos senos simultáneamente, o dos estudios de tiempo diferente) y se enfoca en la diferencia entre ambas. Por tanto, su salida primaria es una medida de similitud/diferencia, en lugar de una clasificación absoluta de una única imagen. Consecuencia de esto es que la red siamesa integra a priori el conocimiento de simetría (lo cual es útil en modalidades bilaterales), mientras que ViT opera a partir de una imagen aislada, aunque uno podría incorporar la imagen contralateral concatenándola como entrada, el modelo ViT estándar no compara dos imágenes explícitamente. En resumen, ViT/SegFormer trabajan sobre imágenes individuales con mecanismos globales de atención, y el siamés trabaja sobre pares de imágenes con mecanismos de comparación directa.

- Requerimientos de datos y entrenamiento: Las redes siamesas han probado ser efectivas en escenarios de datos escasos, ya que pueden aprovechar mejor cada caso al considerarlo como múltiples pares similar/diferente (aumentando implícitamente las muestras de entrenamiento combinando distintas parejas)[53]. Además, su formulación tipo *few-shot* permite aprender a distinguir patrones anómalos con relativamente pocos ejemplos positivos, siempre que existan sus contrapartes negativas (simétricas sanas). Por el contrario, los modelos ViT/SegFormer tienden a requerir más datos etiquetados para entrenar desde cero debido a su alta complejidad. En termografía mamaria, donde las bases de datos suelen ser reducidas (decenas a pocos cientos de pacientes), un ViT sin pre-entrenamiento probablemente no alcanzaría buen rendimiento. Sin embargo, con pre-entrenamiento y/o datos sintéticos adicionales, esta brecha se cierra. De hecho, con la disponibilidad de modelos pre-entrenados, estudios han mostrado que un ViT puede igualar a CNNs entrenadas con pocos datos[54]. Aun así, se podría decir que en igualdad de condiciones (mismo dataset limitado, sin transferencia), el enfoque siamés lograría mejor generalización con pocos datos que un ViT convencional, al menos para la tarea de detección binaria, justamente

---

[51] Garia, L. S., & Hariharan, M. (2023). *Vision Transformers for Breast Cancer Classification from Thermal Images*. En Robotics, Control and Computer Vision (Lecture Notes in Electrical Engineering, vol. 1009), pp. 177–185. DOI: 10.1007/978-981-99-0236-1_13

[52] Dey, A., & Rajan, S. (2024). Thermography-based Breast Abnormality Detection using Siamese Network. *CMBES Proceedings*, *46*. Retrieved from https://proceedings.cmbes.ca/index.php/proceedings/article/view/1186

[53] Ídem.

[54] Garia, L.S., Hariharan, M. (2023). Vision Transformers for Breast Cancer Classification from Thermal Images. In: Muthusamy, H., Botzheim, J., Nayak, R. (eds) Robotics, Control and Computer Vision. Lecture Notes in Electrical Engineering, vol 1009. Springer, Singapore. https://doi.org/10.1007/978-981-99-0236-1_13

porque incorpora conocimiento estructural (simetría) que el ViT tendría que aprender por sí solo.

- Capacidades de clasificación y segmentación: Si el objetivo es únicamente clasificar la presencia o tipo de tumor, ambos enfoques pueden hacerlo, pero de formas distintas. Un ViT típicamente produciría una clasificación multiclase o binaria directa por imagen (por ej. normal vs tumor; o incluso clasificar entre benigno vs maligno si se le entrena para ello). Las redes siamesas, en cambio, se entrenan para diferenciar clases a través de comparación; en el caso de termogramas, esencialmente resuelven una clasificación binaria (simétrico vs asimétrico) que se correlaciona con *ausencia o presencia de anormalidad*. No son naturalmente aptas para multi-clasificación de más de dos estados a menos que se diseñe una extensión (por ejemplo, un siamés por cada par de clases, lo cual complica el sistema). En cuanto a segmentación, la diferencia es grande: el enfoque ViT+SegFormer puede proveer segmentaciones de la lesión de forma end-to-end, entrenado con máscaras pixeladas de los tumores para aprender a marcarlos[55]. SegFormer y similares entregan un mapa donde cada píxel es clasificado (lesión vs fondo), posibilitando delinear con precisión la región caliente del tumor. En cambio, una arquitectura siamesa pura no genera un mapa de segmentación. Para segmentar con ayuda de la red siamesa, habría que recurrir a métodos híbridos, por ejemplo: localizar las regiones más disímiles entre ambas mamas quizás comparando parches correspondientes de la imagen izquierda y derecha. Pero esto no ha sido reportado ampliamente en la literatura. En la práctica, la mayoría de los trabajos con termografía siamesa solo emiten un resultado de detección (p.ej. “se detectó anomalía en esta paciente”)[56] y eventualmente podrían señalar "cuál mama es la anormal" basado en cuál parte contribuyó más al desbalance. Por tanto, para segmentar tumores o zonas específicas, es más conveniente un SegFormer/U-Net entrenado con anotaciones, ya que entrega directamente la máscara (por ejemplo, un estudio reciente usó U-Net para segmentar regiones de interés térmicas con IoU de 0.96 [57]). En resumen, si la tarea incluye segmentación detallada, el enfoque con SegFormer es claramente más apropiado que intentar adaptar una red siamesa, que está orientada principalmente a clasificación por comparación.

---

[55] Sourget, T., Hasany, S.N., Mériaudeau, F., Petitjean, C. (2024). Can SegFormer be a True Competitor to U-Net for Medical Image Segmentation?. In: Waiter, G., Lambrou, T., Leontidis, G., Oren, N., Morris, T., Gordon, S. (eds) Medical Image Understanding and Analysis. MIUA 2023. Lecture Notes in Computer Science, vol 14122. Springer, Cham. https://doi.org/10.1007/978-3-031-48593-0_8

[56] Dey, A., & Rajan, S. (2024). Thermography-based Breast Abnormality Detection using Siamese Network. *CMBES Proceedings*, *46*. Retrieved from https://proceedings.cmbes.ca/index.php/proceedings/article/view/1186

[57] Gualán, R., Jiménez-Gaona, Y., Castillo, D. P., Rodríguez-Alvarez, M., & Lakshminarayanan, V. (2024, October). *Breast thermographic image augmentation using generative adversarial networks (GANs)* (pp. 86–99). https://doi.org/10.1007/978-3-031-75431-9_6

- Rendimiento reportado: Al comparar métricas obtenidas en estudios representativos, vemos que ambos enfoques pueden lograr buenos resultados, aunque bajo condiciones distintas. En termografía mamaria, la red siamesa de Dey (2024) obtuvo ~81% de exactitud en detectar anomalías, trabajando con un conjunto de datos público limitado. Por otro lado, enfoques con CNN tradicionales (sin comparación bilateral) han logrado incluso mayores aciertos cuando el volumen de datos o aumentos lo permitía; por ejemplo, un método basado en texturas y simetría bilateral (sin deep learning) reportó 96% de acierto en termogramas (Dey, 2024), y una combinación de CNN ligera (ResNet34) con un clasificador SVM alcanzó 99.6% en la base DMR-IR[58]. Este último resultado extremadamente alto posiblemente se debe a un entrenamiento potenciado con técnicas de aumento de datos y selección de características, aunque los autores advierten sobre la necesidad de validar la generalización del modelo fuera de esa muestra. En cuanto a Transformers, Garia *et al*. no proporcionan números exactos en su resumen, pero indican que sus modelos ViT lograron desempeños equivalentes o superiores a CNN en la clasificación de termogramas[59] (las CNN típicamente rondaban 90%+ en ese dataset, lo que sugiere que el ViT estuvo en ese rango). En otras modalidades, los Transformers también han mostrado excelente rendimiento: por ejemplo, en mamografía digital, Zhu *et al*. (2021) reportan AUC ~0.98 usando un ViT híbrido para detección de cáncer[60], compitiendo con los mejores CNN. Respecto a segmentación, el estudio comparativo de SegFormer vs U-Net encontró que SegFormer podía igualar la precisión de U-Net en múltiples tareas, con diferencias marginales en los índices de Dice e IoU[61]. Es decir, no hubo pérdida de rendimiento al usar la arquitectura Transformer, e incluso se sugirió considerarla seriamente en futuros desarrollos de segmentación médica. En cambio, una red siamesa por sí sola no tiene resultados publicados en segmentación de termogramas, ya que no produce esa salida directamente. Tomando todo esto en conjunto, para tareas puramente de clasificación/detección con datos muy limitados, las siamesas han probado su eficacia (81–92% de exactitud en casos de uso, con la ventaja de incorporar

[58] Rai, H. M., Yoo, J., Agarwal, S., & Agarwal, N. (2025). LightweightUNet: Multimodal Deep Learning with GAN-Augmented Imaging Data for Efficient Breast Cancer Detection. *Bioengineering (Basel, Switzerland)*, *12*(1), 73. https://doi.org/10.3390/bioengineering12010073

[59] Garia, L.S., Hariharan, M. (2023). Vision Transformers for Breast Cancer Classification from Thermal Images. In: Muthusamy, H., Botzheim, J., Nayak, R. (eds) Robotics, Control and Computer Vision. Lecture Notes in Electrical Engineering, vol 1009. Springer, Singapore. https://doi.org/10.1007/978-981-99-0236-1_13

[60] Ídem.

[61] Sourget, T., Hasany, S.N., Mériaudeau, F., Petitjean, C. (2024). Can SegFormer be a True Competitor to U-Net for Medical Image Segmentation?. In: Waiter, G., Lambrou, T., Leontidis, G., Oren, N., Morris, T., Gordon, S. (eds) Medical Image Understanding and Analysis. MIUA 2023. Lecture Notes in Computer Science, vol 14122. Springer, Cham. https://doi.org/10.1007/978-3-031-48593-0_8

conocimiento humano como la simetría)[62]. Sin embargo, con conjuntos de datos más amplios o aumentados, los Transformers/CNN tienden a sobresalir en rendimiento absoluto, logrando 95–99% en algunos estudios[63]. Además, solo los enfoques tipo ViT+SegFormer permiten abordar detección y segmentación integradas, lo que puede aportar información más rica en un contexto clínico (no solo saber si hay anomalía, sino dónde está).

[62] Dey, A., & Rajan, S. (2024). Thermography-based Breast Abnormality Detection using Siamese Network. *CMBES Proceedings*, *46*. Retrieved from https://proceedings.cmbes.ca/index.php/proceedings/article/view/1186

[63] Rai, H. M., Yoo, J., Agarwal, S., & Agarwal, N. (2025). LightweightUNet: Multimodal Deep Learning with GAN-Augmented Imaging Data for Efficient Breast Cancer Detection. *Bioengineering (Basel, Switzerland)*, *12*(1), 73. https://doi.org/10.3390/bioengineering12010073

12. **Referencias**

1) Abimouloud, M.L., Bensid, K., Elleuch, M., Aiadi, O., Kherallah, M. (2024). Mammography Breast Cancer Classification Using Vision Transformers. In: Abraham, A., Bajaj, A., Hanne, T., Siarry, P. (eds) *Intelligent Systems Design and Applications. ISDA 2023. Lecture Notes in Networks and Systems, vol 1046.* Springer, Cham. https://doi.org/10.1007/978-3-031-64813-7_44
2) Alshehri, A., & AlSaeed, D. (2023). Breast Cancer Diagnosis in Thermography Using Pre-Trained VGG16 with Deep Attention Mechanisms. *Symmetry, 15(3),* 582. https://doi.org/10.3390/sym15030582
3) Alzahem, A., Boulila, W., Koubaa, A. et al. (2023). Improving satellite image classification accuracy using GAN-based data augmentation and vision transformers. *Earth Sci Inform, 16,* 4169–4186. https://doi.org/10.1007/s12145-023-01153-x
4) Bai, J., Jin, A., Wang, T., Yang, C., & Nabavi, S. (2022). Feature fusion Siamese network for breast cancer detection comparing current and prior mammograms. *Medical physics, 49(6),* 3654–3669. https://doi.org/10.1002/mp.15598
5) Becker, A. S., Jendele, L., Skopek, O., Berger, N., Ghafoor, S., Marcon, M., & Konukoglu, E. (2019). Injecting and removing suspicious features in breast imaging with CycleGAN: A pilot study of automated adversarial attacks using neural networks on small images. *European journal of radiology, 120,* 108649. https://doi.org/10.1016/j.ejrad.2019.108649
6) Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End object detection with transformers. *arXiv.* https://arxiv.org/abs/2005.12872
7) Chenqian Li, Jun Liu, Jinshan Tang. (2024). Simultaneous segmentation and classification of colon cancer polyp images using a dual branch multi-task learning network[J]. *Mathematical Biosciences and Engineering, 21(2),* 2024–2049. doi: 10.3934/mbe.2024090
8) Dey, A., & Rajan, S. (2024). Thermography-based Breast Abnormality Detection using Siamese Network. *CMBES Proceedings, 46.* Retrieved from https://proceedings.cmbes.ca/index.php/proceedings/article/view/1186
9) Ervural, S., & Ceylan, M. (2022). Classification of neonatal diseases with limited thermal image data. *Multimedia Tools and Applications, 81,* 1–29. https://doi.org/10.1007/s11042-021-11391-0
10) Garia, L.S., & Hariharan, M. (2023). Vision Transformers for Breast Cancer Classification from Thermal Images. In: Muthusamy, H., Botzheim, J., Nayak, R. (eds) *Robotics, Control and Computer Vision. Lecture Notes in Electrical*

Engineering, vol 1009. Springer, Singapore. https://doi.org/10.1007/978-981-99-0236-1_13

11) Gualán, R., Jiménez-Gaona, Y., Castillo, D. P., Rodríguez-Alvarez, M., & Lakshminarayanan, V. (2024, October). Breast thermographic image augmentation using generative adversarial networks (GANs) (pp. 86–99). https://doi.org/10.1007/978-3-031-75431-9_6
12) Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., & Kleesiek, J. (2024). CellViT: Vision Transformers for precise cell segmentation and classification. *Medical image analysis, 94,* 103143. https://doi.org/10.1016/j.media.2024.103143
13) Jiménez-Gaona, Y., Carrión-Figueroa, D., Lakshminarayanan, V., Rodríguez-Álvarez, M.J. (2024). Gan-based data augmentation to improve breast ultrasound and mammography mass classification, *Biomedical Signal Processing and Control, 94,* 106255. ISSN 1746-8094. https://doi.org/10.1016/j.bspc.2024.106255
14) Kakileti, S. T., Dalmia, A., & Manjunath, G. (2019, June). Exploring deep learning networks for tumour segmentation in infrared images. *Quantitative InfraRed Thermography Journal, 17,* 1–16. https://doi.org/10.1080/17686733.2019.1619355
15) M. A. Rahman, A. Joy, A. T. Abir and T. Shimamura (2024). "Unleashing the Power of Open-Source Transformers in Medical Imaging: Insights from a Brain" *International Journal of Advanced Computer Science and Applications (IJACSA), 15(7)*. http://dx.doi.org/10.14569/IJACSA.2024.01507126
16) Mahoro E., Akhloufi M.A. (2022). Breast cancer classification on thermograms using deep CNN and transformers. *Quant. Infrared Thermogr. J.*, 1–20. doi: 10.1080/17686733.2022.2129135
17) Mahoro, E., & Akhloufi, M. A. (2022). Applying Deep Learning for Breast Cancer Detection in Radiology. *Current oncology (Toronto, Ont.), 29(11),* 8767–8793. https://doi.org/10.3390/curroncol29110690
18) Rai, H. M., Yoo, J., Agarwal, S., & Agarwal, N. (2025). LightweightUNet: Multimodal Deep Learning with GAN-Augmented Imaging Data for Efficient Breast Cancer Detection. *Bioengineering (Basel, Switzerland), 12(1),* 73. https://doi.org/10.3390/bioengineering12010073
19) Sourget, T., Hasany, S.N., Mériaudeau, F., & Petitjean, C. (2024). Can SegFormer be a True Competitor to U-Net for Medical Image Segmentation?. In: Waiter, G., Lambrou, T., Leontidis, G., Oren, N., Morris, T., & Gordon, S. (eds) *Medical Image Understanding and Analysis. MIUA 2023. Lecture Notes in Computer Science, vol 14122*. Springer, Cham. https://doi.org/10.1007/978-3-031-48593-0_8
20) Tagnamas, J., Ramadan, H., Yahyaouy, A. et al. (2024). Multi-task approach based on combined CNN-transformer for efficient segmentation and

classification of breast tumors in ultrasound images. *Vis. Comput. Ind. Biomed. Art, 7,* 2. https://doi.org/10.1186/s42492-024-00155-w

21) Tanimola, O., Shobayo, O., Popoola, O., & Okoyeigbo, O. (2024). Breast Cancer Classification Using Fine-Tuned SWIN Transformer Model on Mammographic Images. *Analytics, 3(4),* 461–475. https://doi.org/10.3390/analytics3040026

22) Tummala, S., Kim, J., & Kadry, S. (2022). BreaST-Net: Multi-Class Classification of Breast Cancer from Histopathological Images Using Ensemble of Swin Transformers. *Mathematics, 10(21),* 4109. https://doi.org/10.3390/math10214109

23) Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *arXiv.* https://arxiv.org/abs/2105.15203

24) Yurtsever, M. M. E., Atay, Y., Arslan, B., & Sagiroglu, S. (2024). Development of brain tumor radiogenomic classification using GAN-based augmentation of MRI slices in the newly released gazi brains dataset. *BMC medical informatics and decision making, 24(1),* 285. https://doi.org/10.1186/s12911-024-02699-6

25) Zhu, C., Chai, X., Xiao, Y., Liu, X., Zhang, R., Yang, Z., & Wang, Z. (2024). Swin-Net: A Swin-Transformer-Based Network Combing with Multi-Scale Features for Segmentation of Breast Tumor Ultrasound Images. *Diagnostics (Basel, Switzerland), 14(3),* 269. https://doi.org/10.3390/diagnostics14030269