CEA Comité Español de Automática
XLVII JORNADAS DE AUTOMÁTICA CÓRDOBA 2026
UNIVERSIDADE DA CORUÑA

Jornadas de Automática

# Arquitectura multimodal para el análisis interpretable de imágenes biomédicas sintéticas

Sergio Lidon-Calvo*, Marina Poveda-Pérez, Marta Nadal-Herraiz, Juliana Manrique-Cordoba, José María Sabater-Navarro
*Unidad de Investigación en Robótica Médica, Universidad Miguel Hernández, Avenida de la Universidad, s/n, 03202, Elche, España.*

**To cite this article:** Lidon-Calvo, S., Poveda-Perez, M., Nadal-Herraiz, M., Manrique-Cordoba, J., Sabater-Navarro, J.M. 2026. Multimodal architecture for interpretable analysis of synthetic biomedical images. Jornadas de Automática, 47. https://doi.org/10.17979/ja-cea.2026.47.13677

**Resumen**

Los modelos de Inteligencia Artificial han demostrado un rendimiento notable, aunque su reducida interpretabilidad plantea un reto en ámbitos donde la comprensión de las decisiones es crucial. Este desafío presenta importancia real en ámbitos clínicos, como en el análisis del cáncer de mama, enfermedad caracterizada por la elevada variabilidad en sus manifestaciones. A fin de abordar dicha limitación, se propone un enfoque multimodal en el que se aplica un conjunto de datos simplificado de origen sintético, el cual combina imágenes junto a un vector de datos adicionales, los cuales guían el aprendizaje del modelo. Para el análisis interpretable, se aplica la herramienta SHAP para evaluar las contribuciones de ambas entradas, concretamente la imagen y las variables del vector. Los resultados evidencian la viabilidad del enfoque propuesto, proponiendo como líneas futuras la optimización del modelo para la disminución de errores de clasificación y la mejora de detección de casos complejos, así como su validación mediante datos reales.

*Palabras clave:* Formulación de modelos, Identificación y validación, Soporte y control de decisiones, Modelado, simulación y visualización de sistemas biomédicos, Cuantificación de parámetros fisiológicos para diagnóstico y evaluación del tratamiento.

**Multimodal architecture for interpretable analysis of synthetic biomedical images**

**Abstract**

Artificial Intelligence models have demonstrated remarkable performance, although their limited interpretability poses a challenge in fields where understanding the decisions made by the model is crucial. This challenge is particularly significant in clinical settings, such as in the analysis of breast cancer, a disease characterized by high variability in its manifestations. To address this limitation, we propose a multimodal approach that uses a simplified synthetic dataset combining images with a vector of additional data to guide the model's learning. For interpretable analysis, we apply the SHAP tool to evaluate the contributions of both inputs, specifically, the image and the vector variables. The results demonstrate the feasibility of the proposed approach, suggesting future directions such as optimizing the model to reduce classification errors and improve the detection of complex cases, as well as validating it using real-world data.

*Keywords:* Model formulation, Identification and validation, Decision support and control, Biomedical system modeling, simulation and visualization, Quantification of physiological parameters for diagnosis and treatment assessment.

## 1. Introducción

En los últimos años, el desarrollo de la Inteligencia Artificial, en concreto, el aprendizaje profundo, ha permitido el avance notable en diversas áreas, entre las que cabe destacar la visión por computador. Los modelos basados en redes neuronales artificiales han demostrado rendimientos destacados en tareas de clasificación y segmentación de imágenes en diversas áreas, impulsado por el desarrollo de novedosos paradigmas, el aumento de recursos computacionales y la acce-

*Autor para correspondencia: slidon@umh.es

sibilidad a conjuntos de datos de gran escala, permitiendo el entrenamiento de redes más complejas y generalizables.

A pesar de ello, dicho enfoque presenta un conjunto de inconvenientes, entre los que cabe señalar la carencia de transparencia explicativa y de trazabilidad en la toma de decisiones de los resultados obtenidos, denominado *Black Box Problem*. Esta limitación radica en el incremento en la complejidad de los modelos basados en redes neuronales, a los cuales, con el propósito de optimizar la precisión y el rendimiento, se les implementa mayor cantidad de capas y, por tanto, se produce un incremento de la cantidad de parámetros entrenables, lo que dificulta la comprensión de su funcionamiento interno. Dicha opacidad obstaculiza la detección de errores, sesgos o correlaciones no deseadas (Ovalle, 2026).

La creciente incorporación de dichas técnicas de Inteligencia Artificial en sectores relevantes, tales como las finanzas, la educación o la sanidad, ha puesto en evidencia la necesidad de implementar modelos y técnicas que permitan la transparencia e interpretabilidad de estos. Ante este desafío, se ha desarrollado el enfoque de la Inteligencia Artificial Explicable (xAI), con el propósito de brindar modelos más confiables, sin la necesidad de disminuir el rendimiento (Ovalle, 2026).

En el contexto de xAI, la interpretabilidad se puede abordar de manera intrínseca, mediante modelos autoexplicativos como los árboles de decisión, o de forma post-hoc, a partir de técnicas que permitan comprender el comportamiento del modelo y la obtención de los resultados tras el proceso de entrenamiento. En lo que respecta al enfoque post-hoc, cabe destacar técnicas como SHAP (SHapley Additive exPlanations) (Lundberg, 2018) y LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016), los cuales proporcionan explicaciones locales en el caso de LIME y tanto locales como globales en el caso de SHAP (Pezzini, 2024).

La interpretabilidad de la Inteligencia Artificial resulta fundamental en el ámbito clínico, particularmente en patologías de elevada complejidad de diagnóstico mediante imagenología médica, como es el cáncer de mama, el tumor maligno más frecuente entre las mujeres, debido a que sus características no son exclusivas. Dicha enfermedad presenta diversas características, tales como la presencia de una masa de morfología irregular con orientación no paralela a la piel, engrosamiento de la dermis y retracción localizada de la zona mamaria en consecuencia a presencia de tumor. El diagnóstico consiste en la combinación de varias modalidades de imagen, entre las que cabe destacar la mamografía o la ecografía (Smithuis et al., 2026; Soler, 2017).

En este caso, existe una amplia literatura basada en técnicas de mapas de calor, tales como SHAP o Grad-CAM (Selvaraju et al., 2017), con los cuales es posible comprender qué píxeles de la imagen son más influyentes en la decisión final. Sin embargo, es imprescindible comprender más allá de la clasificación en benigno o maligno, considerando tanto las áreas visibles de la imagen cómo características adicionales, como la morfología o el tamaño del tumor. En este contexto, se está llevando a cabo el estudio de la aplicación de modelos multimodales, los cuales permiten integrar diversas fuentes de datos, con el objetivo de incrementar la precisión y explicabilidad diagnóstica del cáncer de mama en comparación con enfoques unimodales tradicionales (Lundberg, 2024; Alzamil et al., 2025).

Por ello, el presente estudio preliminar propone un enfoque multimodal explicable aplicado al análisis de cáncer de mama, orientado a favorecer tanto la confianza de los sanitarios en los resultados obtenidos como su posible futura adopción en entornos clínicos reales. Se propone el desarrollo de un conjunto de datos sintético controlado que modela signos morfológicos simples asociados al cáncer de mama, junto a un vector de datos complementarios. Dicho conjunto se emplea en un modelo multimodal, donde la integración de las modalidades se lleva a cabo mediante una fusión intermedia, a partir de un mecanismo de compuertas, el cual permite la modulación entre las modalidades. El modelo es evaluado mediante la herramienta SHAP, con el fin de analizar la contribución a la predicción de cada entrada y de las características individuales, tanto a nivel global como local.

## 2. Materiales y métodos

### *2.1. Conjunto de datos*

El cáncer de mama es una enfermedad que puede manifestarse a través de un conjunto de síntomas físicos. Entre los más relevantes cabe destacar masas con márgenes irregulares debido a la infiltración del tejido subyacente, sombras acústicas posteriores que reflejan la disminución de la señal ultrasonográfica y alteraciones cutáneas, como la depresión o el engrosamiento de la piel, vinculadas a la fijación de la masa en tejido subcutáneo o los ligamentos de Cooper.

Con el propósito de analizar el fenómeno relacionado con la depresión cutánea en dicha afección, se ha generado un conjunto de datos sintéticos para su simulación, compuesto por imágenes en escala de grises de 254 x 254 píxeles. Dichas imagenes han sido replicadas en tres canales, facilitando la extensión a futuros escenarios en los que se haga uso de imágenes RGB reales. La generación de imágenes comienza con la construcción de un fondo en escalas grises al que se le aplica diversas variaciones de intensidad con el objetivo de evitar una apariencia completamente uniforme.

Posteriormente, se definen de manera aleatoria tanto figuras geométricas, tales como círculos y cuadrados, así como variantes con márgenes irregulares, como diversos tipos de líneas, que pueden ser rectas o presentar curvas, las cuales simulan la textura cutánea del paciente. Los elementos incorporados en las imágenes se caracterizan por valores de intensidad asignados de manera aleatoria en el intervalo [0, 255]. Asimismo, se aplica un proceso de desenfoque con el fin de suavizar los bordes, a efectos de enriquecer la complejidad del conjunto de datos (Figura 1).

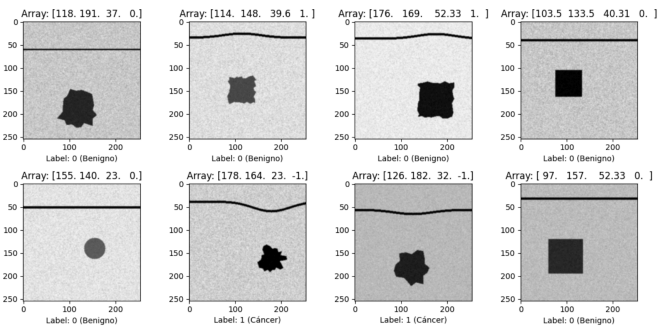



Figura 1: Ejemplo del conjunto de datos.

De manera adicional, se genera un vector formado por información espacial y morfológica (ver Ecuación 1) . En concreto, se incluyen los valores de las coordenadas del eje X e Y del centro de la figura y una medida de escala, expresada como el radio en el caso de los círculos y la mitad de la diagonal para los cuadrados. Además, se incorpora un valor categórico relacionado con el tipo de línea que se representa en la figura. La relevancia de esta clasificación se fundamenta en su interés clínico, puesto que la retracción de los ligamentos de Cooper origina una tensión en la piel, dando lugar a curvaturas visibles en esta, consideradas como un signo indirecto de una lesión tumoral. Dicho valor es codificado como 0 si es una línea recta, 1 en el caso de curvaturas ascendentes y -1 en líneas descendientes. De este modo, la asociación entre la imagen y el array permite la formación de un conjunto de datos multimodal.

$$vector = [centro_X, centro_Y, radio/mitad\_diagonal, tipo\_linea] \quad (1)$$

Si bien el vector de caracteristicas presentado en esta contribución guarda relación con la información de la imagen, el objetivo de este trabajo es presentar la formulación multidimensional. En futuros trabajos se pretende utilizar esta arquitectura cuando la información contenida en el vector de características es adicional a la contenida en la imagen y no puede ser inferida de ésta.

Una vez generado el conjunto de datos, se define un protocolo de asignación de etiquetas en dicho conjunto. Se asigna la etiqueta *Cáncer (1)* a aquellas imágenes que presentan una línea con una onda hacia abajo pronunciada combinada con un círculo irregular. En cambio, la etiqueta *Benigno (0)* corresponde a configuraciones que contienen cuadrados, cuadrados irregulares y círculos, así como líneas rectas o líneas curvas ascendentes o descendentes de menor intensidad.

### 2.2. Modelo multimodal

El modelo desarrollado en el presente estudio está compuesto por dos entradas de datos, una imagen y un array de datos. Ambas entradas son procesadas mediante *encoders* independientes entre sí, formados por capas convolucionales y de agrupación, las cuales se encargan de la reducción del tamaño de los valores de entrada y la cantidad de parámetros, permitiendo la extracción de características representativas para su posterior integración en el modelo.

Tras la extracción de características, se integran capas residuales con el fin de abordar el problema de desvanecimiento del gradiente. A medida que se incrementa la profundidad de la red, la propagación hacia atrás, *backpropagation* puede producir la disminución significativa de los gradientes, dificultando el ajuste de los pesos en el proceso de entrenamiento del modelo. El gradiente de la función de pérdida, calculado en el proceso de retropropagación, indica la dirección del incremento del error, ajustando los parámetros en la dirección contraria, con el propósito de disminuir la función de pérdida, favoreciendo la convergencia del modelo.

Las capas residuales permiten que la entrada de una capa se combine directamente a su salida tras pasar por dos o más transformaciones intermedias. Este enfoque permite que la red aprenda únicamente variaciones relevantes, simplificando la optimización. Además, esta categoría de conexiones facilita la transmisión del gradiente durante el entrenamiento, atenuando las desventajas relacionadas con el desvanecimiento en redes profundas.

En relación con la integración de datos en modelos multimodales, se identifican diversas estrategias de fusión. En el presente trabajo. En este estudio se aplica una fusión intermedia, caracterizada por la extracción de características significativas de cada modalidad para su posterior integración previa a la toma de decisiones. Este enfoque permite el aprendizaje de relaciones complejas entre modalidades manteniendo el procesamiento específico para cada tipo de dato de entrada.

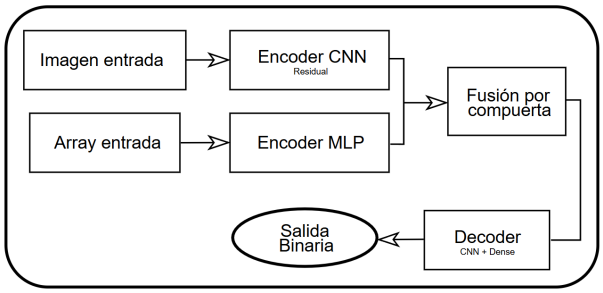



Figura 2: Esquema de la arquitectura de la red.

Concretamente, se aplica una estrategia de fusión intermedia residual con compuerta (Stahlschmidt et al., 2022; Perez et al., 2018). La información visual es modulada por la información textual mediante multiplicación, operando como modulación condicionada. Posteriormente, se adoptan conexiones residuales con el propósito principal de conservar el gradiente en el proceso de propagación, facilitando la conservación del contexto base al integrar de manera aditiva los datos extraídos de la interacción entre las modalidades, resultando en una representación latente más robusta y adaptativa (Figura 2).

### 2.3. Herramienta SHAP

La herramienta SHAP (*SHapley Additive exPlanations*) es un enfoque que permite la interpretabilidad de modelos de Inteligencia Artificial, incluyendo paradigmas de Machine Learning (ML) como de Deep Learning (DL), cuantificando la relevancia de las contribuciones de las variables de entrada. Su aplicabilidad abarca diversos tipos de información, tales como imágenes, señales o datos tabulares, permitiendo su aplicación tanto a nivel global como local (Lundberg, 2024; Alzamil et al., 2025)

El funcionamiento de la herramienta SHAP se fundamenta en la selección de un conjunto de datos de referencia como punto de inicio en la interpretación. El conjunto de referencia seleccionado es un subconjunto de los datos de entrenamiento de igual modo en datos tabulares como imágenes. La relevancia de cada píxel se estima mediante el reemplazo gradual de los píxeles de la imagen original por aquellos correspondientes al conjunto de referencia (Lundberg, 2024; Alzamil et al., 2025).

Una vez definidos los datos de contraste, se procede a la construcción del elemento *Explicador*, el cual es el componente encargado de cuantificar las contribuciones de las variables. Dicho elemento recibe como entrada el modelo entrenado junto al conjunto de referencia y describe cómo la variación

de los datos de entrada afecta a la salida. De acuerdo con la arquitectura del modelo, se reconocen diversos tipos de explicadores, entre los que destacan Tree Explainer, Kernel Explainer y Gradient Explainer. En relación con los modelos basados en redes neuronales artificiales, el uso de Gradient Explainer resulta adecuado, ya que hace uso de la información de los gradientes para estimar las contribuciones.

En el presente caso de estudio, al emplear como conjunto de datos ejemplos sintéticos de morfología de baja complejidad, aplicando únicamente figuras y líneas sencillas, se ha comprobado que la elección del conjunto de referencia influye significativamente en las explicaciones de la herramienta.

La incorporación de un conjunto de datos formado por estructuras definidas de baja complejidad en posiciones variables desencadena el tratamiento de relevancia incorrecto por parte del explicador, generando artefactos procedentes de la comparación con los píxeles de los datos de referencia. Por consiguiente, se establecen como puntos de referencia imágenes de tonalidad gris uniforme, con el propósito de evitar inconsistencias en el proceso de cálculo de las contribuciones de cada uno de los píxeles en las predicciones.

Posteriormente, el explicador se emplea en el cálculo de los valores Shapley sobre un conjunto de muestras específicas. Estos valores representan la influencia marginal de cada característica a la predicción del modelo en una instancia dada.

Esta herramienta asigna a cada píxel de una imagen un valor de SHAP obtenido como el producto entre el gradiente de la salida del modelo respecto al píxel y la diferencia entre el valor del píxel en la imagen original y en el background. Mientras que un valor positivo indica que el píxel contribuye a reforzar la predicción hacia la clase objetivo, un valor negativo denota una influencia contraria (Lundberg, 2024; Alzamil et al., 2025).

La aportación global en las predicciones se obtiene a partir de la suma de los valores absolutos SHAP, mientras que la importancia relativa de cada modalidad mediante la división de importancia de dicha modalidad entre la importancia total, expresándose en porcentaje al multiplicar el valor por 100. Dicho cálculo es posible realizarse sobre muestras en particular.

## 3. Resultados

En la presente sección se exponen los resultados derivados de la aplicación del enfoque predictivo desarrollado en este estudio. El objetivo principal es llevar a cabo el análisis y evaluación del rendimiento del modelo implementado, el cual fue entrenado con un conjunto de datos de entrenamiento formado por 1048 muestras y evaluado mediante un conjunto de prueba de 1048 casos, destinados a la obtención de las métricas.

En el ámbito de la Inteligencia Artificial, las curvas de aprendizaje son una herramienta crucial en la evaluación del desempeño de los modelos, debido a que permiten examinar la evolución del entrenamiento. El contraste entre las diversas curvas en los conjuntos de entrenamiento y prueba permite la identificación de sobreajuste o déficit de ajuste del modelo, proporcionando criterios de optimización de hiperparámetros o la propia arquitectura del modelo.

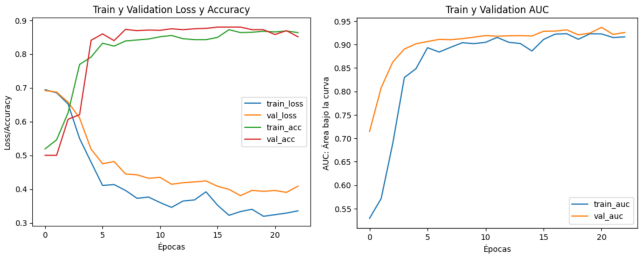



Figura 3: Curvas de entrenamiento y validación de pérdida, precisión y AUC.

En la Figura 3 se muestra la comparación de las curvas de pérdida, precisión y AUC en los conjuntos de entrenamiento y prueba. En primer término, es pertinente señalar que el entrenamiento se detuvo aproximadamente en la época 20, sin haberse completado las 50 épocas establecidas, como resultado del callback *EarlyStopping* al no registrarse mejoras en las pérdidas de validación durante 5 épocas.

Las curvas presentan un descenso significativo en la primera fase del entrenamiento, seguido de la estabilización continua cercana a cero, evidenciando un aprendizaje progresivo. Al tratarse de un conjunto de datos de baja complejidad, el modelo posee de elevada capacidad de aprendizaje de los patrones subyacentes, ajustando los parámetros en un número menor de iteraciones. Una vez expuestas las gráficas relacionadas con el proceso de entrenamiento, se incluye la tabla correspondiente a las métricas del modelo con el objetivo de evaluar su rendimiento de manera precisa.

Tabla 1: Métricas de evaluación del modelo multimodal.

| | Benigno (0) | Cáncer (1) | Global |
|---|---|---|---|
| Precisión | 0.90 | 0.85 | - |
| Exhaustividad | 0.84 | 0.90 | - |
| Puntuación F1 | 0.87 | 0.87 | - |
| Soporte | 524 | 524 | 1048 |
| Precisión | - | - | 0.87 |
| AUC | - | - | 0.87 |
| Pérdida | - | - | -0.001 |

Como se muestra en la Tabla 1, el modelo multimodal alcanza una precisión de 0.87, evidenciando un rendimiento global satisfactorio de predicción. En cuanto a las métricas por clase, la precisión es de 0.90 en la clase 0 y 0.85 en la clase 1, mientras que, en el caso de la Exhaustividad, el cual define la capacidad de identificación de los casos positivos reales, se sitúan los valores en 0.84 y 0.90. La distribución de clases es uniforme, con 524 muestras en cada una, mitigando así la problemática del sesgo de clases. A nivel global, se observa una precisión y un AUC de valor 0.87, junto con una pérdida mínima del -0.001, indicativos de una discriminación correcta entre clases.

Tras el análisis de las métricas obtenidas, es relevante evaluar su comportamiento con mayor detalle mediante la matriz de confusión. Como se refleja en la Figura 4, el modelo alcanza un rendimiento predictivo satisfactorio en ambas clases, indicando una leve ventaja en la clase 1 (Maligno) que cuenta con 51 falsos negativos frente a 85 falsos positivos en la clase 0, lo que evidencia una mayor penalización en la clase benigna, aunque el balance general de los errores es equilibrado.

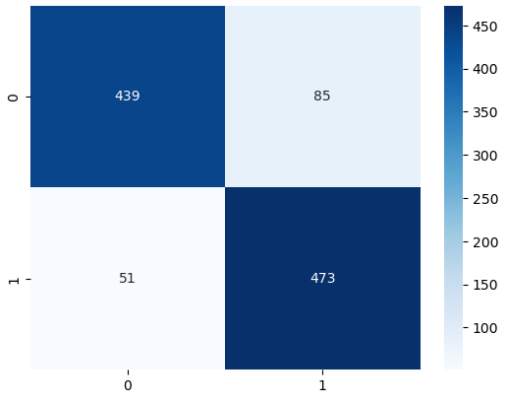
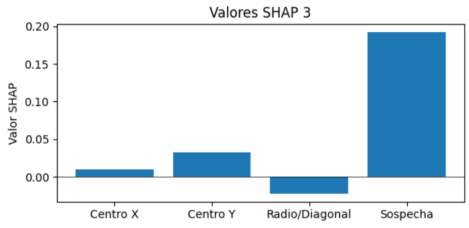
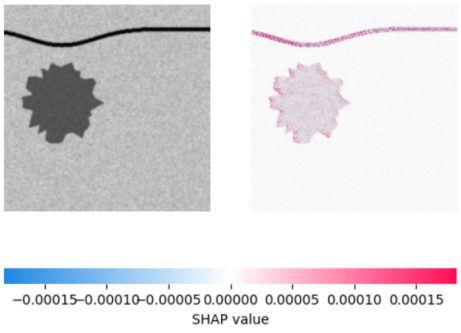



Figura 4: Matriz de confusión del enfoque propuesto.

Una vez validado el rendimiento a partir de las métricas obtenidas y la matriz de confusión, se procede a analizar el comportamiento de cada una de las modalidades de entrada en el proceso de predicción del modelo, permitiendo analizar tanto su significancia a nivel global, como en el orden de características particulares. A nivel global, se evidencia una contribución del 53.81 % por parte de la imagen, mientras que el vector de datos contribuye un 46.19 % en la predicción.




(a) Valores SHAP en las variables tabulares.




(b) Valores SHAP en la imagen.

Figura 5: Valores SHAP en un caso de estudio.

En la Figura 5 se presentan los resultados SHAP en un ejemplo representativo. En cuanto a las variables tabulares (Figura 5(a)), se evidencia que la variable *Sospecha*, es decir, el tipo de linea, constituye la de mayor relevancia, seguida del *Centro Y*. El mapa de valores SHAP de la entrada visual (Figura 5(b)) permite diferenciar regiones con contribuciones positivas (rojas) y negativas (azules) en la predicción.

## 4. Discusión de resultados

El análisis de los resultados obtenidos evidencia un rendimiento equilibrado del modelo propuesto en el presente estudio (Tabla 1). La puntuación F1 presenta valores de 0.87 en ambas clases, junto con un AUC de 0.87, lo que indica un equilibro del modelo y una capacidad adecuada para distinguir entre casos benignos y cancerígenos.

Respecto al análisis diferenciado por clase, se observa mayor precisión en la clase benigna (0.90) respecto con la clase maligna (0.85). Además, la exhaustividad adopta un valor de 0.90 en la clase maligna. Este comportamiento denota una baja tasa de falsos positivos y elevada sensibilidad ante casos positivos, aspecto esencial en el ámbito clínico.

Trascendiendo los indicadores globales, es fundamental la interpretación de los valores SHAP, los cuales permiten comprender las contribuciones de ambas modalidades de entrada, así como de cada característica en la predicción de cada uno de los casos.

Si bien el análisis global demuestra contribuciones equitativas entre modalidades, lo cual es coherente con el hecho de que el vector contiene información que también aparece en la imagen, es relevante analizar casos de manera independiente con el objetivo de evaluar posibles desviaciones en el patrón de relevancia en la predicción entre la imagen y el array de datos (Figura 6).

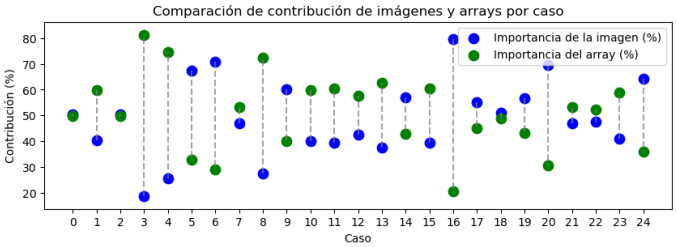



Figura 6: Comparación de contribuciones en 25 casos específicos.

La Figura 6 muestra que, en diversas estancias, predomina la información visual, como sucede en los casos 5, 6 y 16, mientras que en otros casos priman los datos textuales, como se evidencia en los ejemplos 3 y 4. Dicha variación se fundamenta en un mecanismo por el cual el modelo prioriza la información de mayor naturaleza discriminativa, según los indicadores en cada modalidad.

En los escenarios 3 y 4, el descriptor asociado al tipo de línea del array de datos presenta un nivel elevado discriminativo. En el caso 3, correspondiente a un ejemplo de cáncer, compuesto por una línea curva inferior pronunciada y un círculo irregular, permite al modelo tomar una decisión adecuada debido al tipo de línea. En consecuencia, el modelo prioriza la información contenida en el array, reduciendo la necesidad de un análisis exhaustivo de la imagen para la clasificación. En el caso 4, la combinación de un cuadrado y una línea con curvatura ascendente se asocia a datos benignos, lo que deriva en una priorización de los datos del array frente a la información de la imagen.

En contraste, en los casos 5 y 6, la línea presenta una curvatura atenuada que impide la extracción de conclusiones definitivas, ya que genera incertidumbre. En consecuencia, el

modelo depende de la información visual en la toma de decisiones final, donde la morfología del objeto cuadrangular proporciona un indicador más robusto en la interpretación en comparación con los datos tabulares. De manera análoga, a diferencia de otros escenarios en los que la *Sospecha* presenta un comportamiento predecible, en el ejemplo 16 se observa un valor SHAP negativo de dicha variable, a pesar de que el valor -1 indica una curvatura inferior en la línea. La discrepancia entre la morfología de la línea con un cuadrado irregular induce al modelo a verificar la información visual. De este modo, se disminuye la ponderación de la contribución del array, otorgando a los píxeles influencia decisiva en la predicción.

A pesar de ello, la mayor parte de las muestras mantiene un equilibro en la coordinación de ambas modalidades, lo que evidencia un tratamiento multimodal eficaz. La variabilidad de la dependencia de dichas modalidades se asocia a la naturaleza del conjunto de datos y la asignación de etiquetas del mismo, junto a las particularidades de cada instancia, mostrando una integración satisfactoria entre información visual y textual.

Sin embargo, se debe considerar limitaciones. La variabilidad observada en la contribución específica en cada caso supone una evidencia de dependencia en las características de entrada de cada muestra. Por tanto, dicho suceso puede suponer una disminución de la estabilidad en las predicciones en casos de estudio en los que predomina una modalidad en específico.

## 5. Conclusiones y trabajos futuros

Este estudio valida un marco multimodal aplicado en un entorno simplificado, el cual emula ciertos signos identificativos del cáncer de mama. La arquitectura demuestra un rendimiento robusto en tareas de clasificación, mediante la integración de imágenes y datos que no siempre son perceptibles en la imagen, enriqueciendo la capacidad de aprendizaje del modelo.

No obstante, los resultados deben considerarse en un entorno simplificado y controlado, lo que puede limitar su extrapolación a entornos reales. Además, la contribución entre modalidades sugiere un margen de optimización en la fusión multimodal.

Como trabajos futuros, se propone la validación del modelo mediante datos clínicos reales, junto con variables clínicas adicionales. Asimismo, resulta fundamental optimizar las estrategias de clasificación con el propósito principal de mejorar la detección de casos complejos y reducir los errores. El modelo podría beneficiarse de la integración de mecanismos de atención más avanzados y adaptativos, orientados a optimizar el proceso de fusión y robustez en entornos de información multimodal.

## Agradecimientos

Esta investigación ha sido financiada por el Gobierno español —Agencia Estatal de Investigación (AEI)— a través del proyecto PID2022-138206OB-C32, y por la Generalitat Valenciana a través de los proyectos INNVA1/2024/59 y CIPROM2022/16.

## Referencias

Alzamil, D., Alkhamees, B., Hassan, M. M., 2025. A systematic review of multimodal fusion and explainable ai applications in breast cancer diagnosis. Computer Modeling in Engineering & Sciences 145 (3), 2971.

Lundberg, S., 2018. Welcome to the shap documentation. `https://shap.readthedocs.io/en/latest/`, accessed: 2026-05-06.

Lundberg, S., 2024. Shap image interpretation examples. `https://shap.readthedocs.io/en/latest/image_examples.html`, accessed: 2026-03-31.

Ovalle, M. T. O., 2026. Inteligencia artificial explicable (xai) como marco para la transparencia en la ciencia de datos: Interpretabilidad, toma de decisiones ética y responsabilidad algorítmica. Revista Científica de Salud y Desarrollo Humano 7 (1), 258–283.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A., 2018. Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. Vol. 32.

Pezzini, M. C., 2024. Inteligencia artificial explicable: análisis de metodologías y aplicaciones. Ph.D. thesis, Universidad Nacional de La Plata.

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

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626.

Smithuis, R., Wijers, L., Dennert, I., 2026. Ultrasound of the breast. `https://radiologyassistant.nl/breast/ultrasound/ultrasound-of-the-breast`, accessed March 2026.

Soler, J. R., 2017. Herramientas computacionales basadas en imagen de ultrasonografía para diagnóstico médico. Ph.D. thesis, Universidad Miguel Hernández de Elche.

Stahlschmidt, S. R., Ulfenborg, B., Synnergren, J., 2022. Multimodal deep learning for biomedical data fusion: a review. Briefings in bioinformatics 23 (2), bbab569.