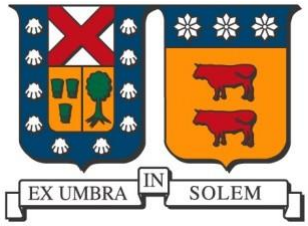
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
VALPARAÍSO - CHILE



"EVALUACIÓN DE INCERTIDUMBRE EN LA SEGMENTACIÓN AUTOMÁTICA DE HÍGADO Y SEGMENTOS DE COUINAUD EN IMÁGENES DE RESONANCIA MAGNÉTICA"

RHAIMANHS RAÚL LABARCA GALLARDO

TESINA PARA OPTAR AL GRADO
DE MAGÍSTER EN TECNOLOGÍAS DE
LA INFORMACIÓN

Profesor Guía: JULIO SOTELO

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☐ Memoria o trabajo de título **X** Tesis de Postgrado

Título del trabajo: Evaluación de incertidumbre en la segmentación automática de hígado y segmentos de Couinaud en imágenes de resonancia magnética

Nombre del candidato(a): Rhaimanhs Raúl Labarca Gallardo

Carrera / Grado: Magíster en Tecnologías de la Información.

Campus: Vitacura **Departamento:** informática

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, **Julio Sotelo**, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

X El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 28-10-2026 **Firma:**



Estudiante o Candidato(a):

Fecha: 28-10-2026 **Firma:**



Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Evaluación de incertidumbre en la segmentación automática de hígado y segmentos de Couinaud en imágenes de resonancia magnética

Rhaimanhs Raúl Labarca Gallardo[1]

Departamento de Informática,
Universidad Técnica Federico Santa María, Santiago, Chile.

rhaimanhs.labarca@sansano.usm.cl

Resumen: La segmentación del hígado y de sus ocho segmentos de Couinaud en resonancia magnética (RM) la realizan manualmente los radiólogos, generando mucha variabilidad, y las soluciones automatizadas no advierten dónde pueden fallar. Es por ello que se propuso evaluar la incertidumbre en tres arquitecturas convolucionales (U-Net 3D, ResEnc M y MedNeXt-B) entrenadas con 197 volúmenes de RM de 73 pacientes. La evaluación se realizó sobre 30 volúmenes utilizando métricas de volumetría, sensibilidad geométrica, estimación de incertidumbre por aumentación en tiempo de prueba y mapas de incertidumbre. Como resultado, el hígado y sus segmentos obtuvieron un Dice de 0,937 y 0,778 respectivamente; las arquitecturas no difirieron en la calidad de la incertidumbre tras la corrección de Holm. La incertidumbre fue nueve veces mayor en los vóxeles erróneos del contorno hepático y tres veces mayor en las fronteras intersegmentarias (AUC-ROC de 0,81 y 0,72). Los volúmenes con mayor incertidumbre fueron los peor segmentados con una correlación de Spearman (ρ) entre -0,61 y -0,74. Los segmentos I, III y VII mostraron mayor variabilidad volumétrica (38,6% frente al 3% del hígado completo). El impacto esperado es reducir el tiempo de revisión de la segmentación, orientando al radiólogo hacia las zonas de mayor incertidumbre.

Palabras Clave: segmentación hepática, resonancia magnética, cuantificación de incertidumbre, segmentos de Couinaud, aumentación del tiempo de prueba.

1 Introducción

La segmentación automática del hígado en resonancia magnética alcanza hoy valores de Dice superiores a 0,92 en el órgano completo [1], [2], pero los modelos entregan su resultado sin indicar cuán confiable es, y esa carencia se agudiza al subdividir el hígado en los ocho segmentos de Couinaud, cuyas fronteras no son visibles en la imagen. Este trabajo aborda ese problema evaluando la incertidumbre de distintas arquitecturas de segmentación, tanto a nivel del hígado completo como de cada segmento.

1.1 Contexto

Para planificar una cirugía de hígado, medir su volumen o seguir la evolución de un tumor, primero es necesario delimitar el órgano en las imágenes y dividirlo en los ocho segmentos descritos por Couinaud [3]. Esta división tiene una razón práctica: cada segmento tiene irrigación sanguínea y drenaje propio [4], de modo que el cirujano puede extirpar uno o varios sin dañar el tejido hepático que queda. En Chile, más de doscientos pacientes esperan un trasplante de hígado y se realizan cerca de ciento setenta trasplantes al año [5]. En ese contexto, medir con precisión el volumen de cada segmento es clave para estimar si la parte del hígado que quedará después de una cirugía será suficiente.

[1] Ingeniero en Computación y Licenciado en Ciencias de la Computación.

Hoy, esta tarea la realiza manualmente un especialista y toma entre treinta y sesenta minutos por estudio, según la experiencia del radiólogo vinculado al proyecto. El mayor desafío no radica en delinear el contorno exterior del hígado, el cual es distinguible, sino en trazar los límites intersegmentarios, los cuales no son visibles de forma directa, y el radiólogo los deduce trazando planos que pasan por ciertas venas del hígado (las venas suprahepáticas y el punto donde se divide la vena porta).

La resonancia magnética (RM) es el examen más usado para estudiar el hígado, porque muestra con gran detalle los tejidos blandos. Sin embargo, tiene tres características que complican la segmentación automática. Primero, los contrastes de la imagen pueden variar según el estudio, el equipo o el centro médico. Segundo, la resolución de la imagen puede contener vóxeles que no son isotrópicos, debido a que la distancia entre cortes es mayor que el detalle dentro de cada corte. De hecho, en la base utilizada, el 59% de los volúmenes tiene vóxeles más de tres veces más largos en una dirección que en las otras. Tercero, hay pocos conjuntos de imágenes ya segmentadas para entrenar algoritmos debido a que la resonancia magnética se ha estudiado menos que la tomografía computarizada (TC) y no existen bases públicas de referencia para dividir el hígado en segmentos [2].

1.2 Definición del problema

Dadas las condiciones descritas, se sigue que un modelo entrenado sobre RM se equivocará con más frecuencia que uno desarrollado en condiciones más favorables. El problema que aborda esta tesina no es el desempeño como tal, sino algo que lo acompaña: estos modelos entregan una segmentación sin decir en qué partes de ella se puede confiar. La salida de un modelo de segmentación es, en cada vóxel, una distribución de probabilidades entre las distintas estructuras anatómicas posibles (como el tejido sano, el fondo o cada segmento hepático). El modelo asigna únicamente la categoría de mayor probabilidad y descarta el resto, entregando un mapa de etiquetas donde todos los vóxeles tienen el mismo aspecto; los que el modelo en su predicción acertó con holgura y los que decidió por un margen mínimo. Al carecer de un indicador de confianza, el radiólogo que recibe esa segmentación no tiene forma de saber dónde le conviene mirar con atención, así que no le queda más que revisarla completa, y con ello se pierde buena parte del beneficio de automatizarla.

Podría pensarse que basta con mostrar la probabilidad que el modelo asignó a cada vóxel. Sin embargo, esta aproximación es insuficiente, porque las redes profundas modernas tienden a estar mal calibradas [6], y declaran una confianza superior a su exactitud real. Una red puede afirmar que un vóxel pertenece al quinto segmento con una probabilidad de 0,99 y equivocarse. No olvidar que dicho valor refleja el nivel de confianza matemática del modelo, no su probabilidad real de acierto. Se necesita entonces un procedimiento externo que estime la confianza de otra manera, y una forma de comprobar si esa estimación se corresponde con los errores reales. El problema quedó formulado en dos preguntas. La primera es, ¿Qué arquitectura produce la estimación de incertidumbre más confiable en segmentación hepática sobre RM?, entendiendo por confiable aquella cuya magnitud coincide efectivamente con la probabilidad de error, es decir, que haya incertidumbre alta donde el modelo se equivoca y baja donde acierta. La segunda es, ¿Qué segmentos de Couinaud concentran la mayor incertidumbre, medida tanto por vóxel como en volumen?, porque el volumen es la unidad en que estos resultados se usan clínicamente.

1.3 Solución Propuesta y Objetivos

Se evaluaron comparativamente tres arquitecturas convolucionales (U-Net 3D, ResEnc M y MedNeXt-B), estimando la incertidumbre mediante aumentación en tiempo de prueba con transformaciones geométricas invertibles. Se valida la fiabilidad de esta estimación a través de métricas independientes de la calidad de la segmentación.

El objetivo general fue evaluar la calidad de la incertidumbre producida por distintas arquitecturas de segmentación hepática en RM, a nivel del hígado completo y de cada segmento de Couinaud.

Los objetivos específicos fueron:

- Construir una base anotada con los ocho segmentos de Couinaud sobre RM, con protocolo documentado y validado por un especialista.
- Construir un pipeline reproducible de preprocesamiento con orientación y espaciado físico homogéneos.
- Entrenar y comparar arquitecturas sobre las tareas de hígado completo y segmentos.
- Evaluar la sensibilidad a rotaciones y escalados conocidos sobre los tres ejes anatómicos.
- Estimar la incertidumbre por vóxel y por segmento, con inversión geométrica de las transformaciones, y expresarla en volumen.
- Cuantificar su calidad mediante el área bajo la curva ROC para detectar el error (AUC-ROC de detección), la razón de entropía entre vóxeles erróneos y correctos, y el error de calibración esperado (ECE).

1.4 Hipótesis

La estimación de incertidumbre por aumentación en tiempo de prueba permite identificar, sin anotación manual previa, las regiones y los estudios en que la delimitación automática de los segmentos de Couinaud es menos confiable, y expresar esa confianza en volumen. Se complementa con tres hipótesis específicas.

H1. Arquitecturas con Dice similar (diferencia < 0,02) presentan diferencias > 0,05 en su AUC-ROC de detección.

H2. El modelo con menor ECE, supera en al menos 0,03 de AUC-ROC de detección al de mayor ECE.

H3. Sobre los ocho segmentos de Couinaud, la razón de entropía entre vóxeles erróneos y correctos supera a 3, y el AUC-ROC de detección queda bajo 0,75. Ambos umbrales fijados son criterio propio de este trabajo.

La validación se apoya en un diseño común a las tres hipótesis. Para garantizar que las diferencias de rendimiento se deban puramente a la arquitectura de las redes y no al proceso, las tres comparten exactamente las mismas condiciones:

- Se entrenan con la misma división de datos y durante la misma cantidad de épocas (250).
- Se evalúan utilizando el mismo código sobre un conjunto de prueba independiente, apartado desde el principio (30 volúmenes obtenidos de 11 pacientes).

Cada hipótesis se evalúa con herramientas específicas:

- **H1 y H2 (Calidad de la incertidumbre):** Se compara cuán bien cada red estima sus propias dudas. Para ello, se rotan o alteran geométricamente las imágenes durante la prueba y se miden los resultados usando el AUC-ROC de detección (que cuantifica si la incertidumbre permite separar los vóxeles correctos de los erróneos) y el error de calibración esperado (que mide la coherencia entre la confianza declarada por el modelo y su exactitud observada). Estas diferencias se validan estadísticamente utilizando la prueba de Wilcoxon pareada volumen por volumen, con corrección de Holm.
- **H3 (Origen del error):** Se busca determinar si el error del modelo deriva de un fallo algorítmico o de la difuminación inherente de los bordes anatómicos en la imagen. Esto se logra analizando en qué zonas de la imagen duda más el modelo (bordes o interior) y calculando qué porcentaje de errores logra advertir por sí solo.

Las pruebas se respaldan con dos controles para evitar sesgos:

- **Control de reversibilidad:** Verifica que rotar una imagen y devolverla a su posición original no introduzca errores artificiales. Se mide como la proporción de vóxeles del volumen de referencia que quedan mal asignados tras este ciclo.
- **Límite superior geométrico:** Establece cuál sería el Dice máximo que obtendría un modelo perfecto bajo cada transformación, es decir, un rendimiento máximo ideal. Se calcula aplicando la

transformación a la segmentación de referencia y revirtiéndola, sin intervención del modelo. Esto permite diferenciar si una degradación en el Dice se debe a un error del modelo o simplemente a la distorsión matemática de manipular la imagen.

1.5 Organización del Informe

El Capítulo 2 presenta el marco teórico, el estado del arte y los trabajos relacionados. El Capítulo 3 describe la base de datos, el protocolo de anotación, las implementaciones desarrolladas y el procedimiento de estimación de incertidumbre. El Capítulo 4 presenta los resultados y su análisis. El Capítulo 5 expone las conclusiones, limitaciones y trabajo futuro.

2 Marco Teórico

2.1 Marco conceptual

Anatomía segmentaria del hígado. La clasificación de Couinaud [3] divide el parénquima hepático[2] en ocho segmentos funcionalmente independientes. Los planos divisorios quedan definidos por las tres venas suprahepáticas en el eje vertical y por un plano transversal imaginario que pasa por la bifurcación de la vena porta (figura 1-A).

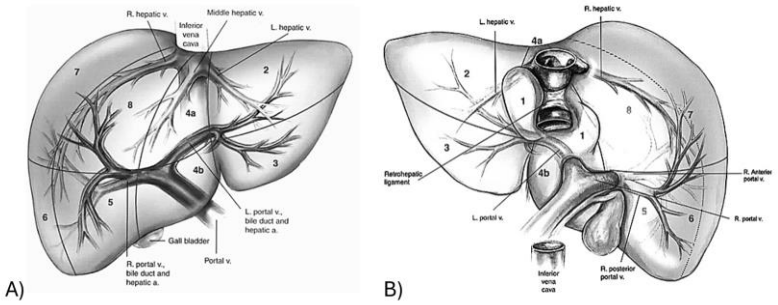


Figura 1. Anatomía vascular y biliar intrahepática. Izquierda vista anterior, derecha vista posterior. Reproducida de Abdel-Mishih y Bloomston [4] (figuras 7 y 8), adaptadas a su vez de Cameron y Sandone [30]

El segmento I presenta una independencia anatómica adicional al drenar directamente a la vena cava inferior [4] (figura 1-B). Esta topología anatómica impone dos restricciones que condicionaron directamente el diseño computacional. Las etiquetas constituyen una partición espacial estricta del hígado, sin huecos ni solapamientos; y los volúmenes intersegmentarios son marcadamente asimétricos (el segmento I es aproximadamente cinco veces más pequeño que sus vecinos). Esta asimetría induce un severo desbalance de clases que afecta tanto en la optimización de la función de pérdida, como en la posterior interpretación de las métricas de rendimiento. Conviene distinguir la definición anatómica pura basada en territorios de irrigación portal (que genera fronteras irregulares y entrelazadas) de la definición radiológica por planos, la cual aproxima superficies divisorias lisas a partir de las referencias vasculares mayores.

Arquitecturas de segmentación. La arquitectura U-Net [7] funciona en dos etapas: primero, reduce la imagen para entender la anatomía general y luego la amplía para recuperar el detalle. Para no perder nitidez al ampliar, usa conexiones de salto (skip connections) que transmiten los bordes y contornos clave (figura 2). Su variante volumétrica (U-Net 3D) [8] permite capturar el contexto tridimensional, aunque a expensas de un consumo de

[2] Parénquima hepático: Masa de tejido funcional y activo del hígado, formado principalmente por hepatocitos, encargado de realizar las tareas metabólicas, de depuración y de síntesis del órgano.

memoria que crece con la tercera potencia de la dimensión lineal del volumen de entrada. Sobre esta estructura base, destacan diversas mejoras: los codificadores residuales, que facilitan el flujo del gradiente durante el entrenamiento (esto evita que el modelo "olvide" o degrade la información); los mecanismos de atención [9], que priorizan las zonas relevantes de la imagen en las conexiones de salto (enfocarse en las regiones anatómicas claves y filtrar el ruido de fondo); y los bloques tipo ConvNeXt integrados en arquitecturas como MedNeXt [10], los cuales emplean convoluciones separables en profundidad para maximizar el campo receptivo por parámetro, es decir, estos bloques amplían el área de visión de la red de forma eficiente sin consumir demasiada memoria. Por su parte, el marco nnU-Net [11] elimina la configuración manual al analizar los datos y definir automáticamente, mediante reglas heurísticas, el preprocesamiento, las dimensiones de los bloques de entrada (patch size), la topología y los hiperparámetros óptimos. Una revisión del año 2024 [12] de este marco mostró que, bajo esquemas de validación rigurosos, muchas de las ventajas de diversas mejoras arquitectónicas recientes se diluyen, dejando a la variante con codificador residual como una línea base sólida, una premisa que los hallazgos empíricos del presente trabajo permitirán discutir.

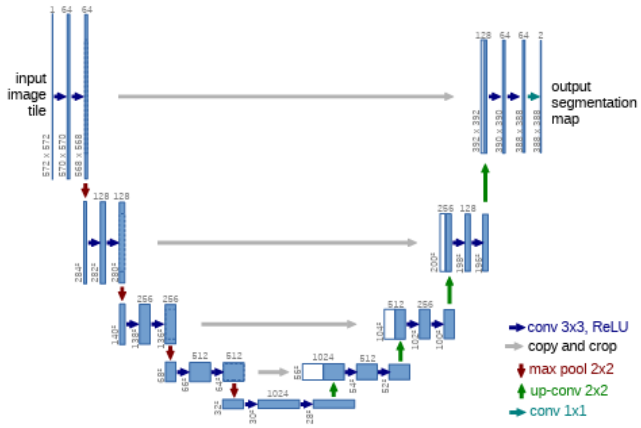


Figura 2. Arquitectura U-Net (ejemplo para imágenes de 32x32 píxeles en la resolución más baja). Figura original de Ronneberger et al. [7]

Cuantificación de incertidumbre. Se distingue entre incertidumbre aleatoria (ruido en los datos) y epistémica (limitaciones o falta de conocimiento del modelo). Para su estimación en imágenes médicas predominan tres métodos. Primero, el método de Monte Carlo dropout [13], que mantiene activas las capas de abandono (dropout) durante la fase de inferencia, aunque su comportamiento depende de la tasa y de dónde se ubica en la red a la capa de dropout (en el codificador, en el embotellamiento o en el decodificador), lo que dificulta una comparación equitativa. Segundo, los conjuntos profundos (*deep ensembles*) [14], que se basan en el entrenamiento de múltiples modelos independientes, con alto desempeño, pero excesivo costo computacional; y, por último, la aumentación en tiempo de prueba (TTA, por *test-time augmentation, en sus siglas en inglés*) [15] que aplica leves modificaciones a la imagen (ej. giros), segmenta cada variante, deshace el cambio y promedia los resultados. Al ser agnóstica a la arquitectura y prescindir de reentrenamiento, la TTA es idónea para comparar modelos en igualdad de condiciones. No obstante, Wang et al. [15] la presentan como estimador de incertidumbre aleatoria; aquí se adopta el procedimiento sin esa lectura, porque las transformaciones geométricas prueban la estabilidad ante el posicionamiento, no la aleatoria ni la epistémica. Además, una red puede repetir el mismo error en todas las variaciones y parecer engañosamente segura.

Inversión de las transformaciones. Aunque las rotaciones geométricas son invertibles, su aplicación sobre mallas discretas de vóxeles exige procesos de interpolación que no lo son. Por ello, revertir un mapa de etiquetas genera errores de asignación en los bordes, inflando artificialmente la incertidumbre en estas zonas críticas. Para evitarlo, se adoptó el enfoque documentado en [16], el cual transforma coordenadas espaciales en lugar de reinterpolar la imagen. Para cada vóxel original, se calcula su posición matemática en el dominio transformado, redondeando al índice entero más cercano y se extrae la etiqueta predicha en dicha posición.

Métricas de evaluación. El coeficiente de Dice mide el solapamiento entre la predicción y la referencia. El error de contorno se cuantifica en milímetros reales mediante la distancia media de superficie (ASSD) y el percentil 95 de la distancia de Hausdorff (HD95) [17]. El error de calibración esperado (ECE) [6] evalúa la discrepancia entre la confianza declarada por el modelo y su exactitud real. Finalmente, la calidad de la incertidumbre se evalúa mediante el área bajo la curva ROC, en adelante AUC-ROC, que corresponde a la probabilidad de que un vóxel mal clasificado registre mayor incertidumbre que uno correcto. Un valor de 0,5 indica que la incertidumbre no aporta información útil (equivalente a lanzar una moneda al aire); por otro lado, un valor de 1,0 significa que el modelo detecta perfectamente sus propios errores. Además, se calcula cómo varía el volumen total estimado por el modelo (mediante la desviación estándar y coeficiente de variación). Estas últimas métricas se ajustan estrictamente al contexto clínico, donde el parámetro de decisión es el volumen anatómico absoluto y no el solapamiento espacial.

2.2 Estado del arte

El rendimiento actual de los algoritmos para la segmentación global del hígado en resonancia magnética (RM) ha alcanzado un alto nivel de madurez, situándose consistentemente con un Dice entre 0,92 y 0,97. Por ejemplo, Zbinden et al. [1] lograron un Dice de 0,936 sobre adquisiciones T1 Vibe Dixon sin contraste (reportando además 0,634 para la vena porta y 0,532 para las venas hepáticas), mientras que Raab et al. [2] alcanzaron 0,97 empleando RM multifase con contraste hepatobiliar (ácido gadolinio etoxibencil-dietilentriaminopentaacético, Gd-EOB-DTPA por sus siglas en inglés), superando incluso el valor de referencia de 0,954 establecido en el desafío CHAOS [18]. En la misma línea, Oh et al. [19] desarrollaron una segmentación hepática automática en fase hepatobiliar orientada a la planificación preoperatoria, y Hänsch et al. [20] mostraron que combinar varios modelos mejoran la segmentación de lesiones hepáticas en RM de fase tardía. Sin embargo, a diferencia de la delineación del hígado completo, la subdivisión anatómica funcional en los ocho segmentos de Couinaud sobre RM presenta escasos antecedentes y un grado de complejidad considerablemente mayor. El referente metodológico más afín al presente trabajo (por modalidad, tarea y volumen de datos) es el estudio de Zbinden et al. [21], quienes entrenaron una red convolucional sobre 200 estudios T1 Vibe Dixon sin contraste, obteniendo un Dice medio de 0,770 (con oscilaciones entre 0,726 para el segmento IVb y 0,810 para el V). Aunque otros autores reportan métricas superiores, sus aproximaciones se apoyan en condiciones considerablemente más favorables. Han et al. [22], por ejemplo, alcanzaron un Dice de 0,920 utilizando imágenes en fase portal y un conjunto de datos de entrenamiento cuatro veces mayor. De manera similar, Miao et al. [23] obtuvieron 0,853 sobre RM contrastada mediante un modelo multitarea que combina la segmentación volumétrica con la detección de siete hitos vasculares, mientras que Nicoli et al. [24] reportaron 0,847 empleando el marco nnU-Net, cifra que asciende a 0,897 al aplicar la misma técnica sobre tomografía computarizada (TC). Una vía alternativa, propuesta por Arya et al. [25], prescinde de la segmentación directa y localiza coordenadas sobre las venas para inferir matemáticamente los planos divisorios; no obstante, su método omite el lóbulo caudado (segmento I), una limitación que esquemas multitarea como el de Miao et al. [23] logran resolver. La brecha de rendimiento observada de forma consistente entre modalidades radica en un principio elemental, en el cual, en la TC con contraste las venas que definen las fronteras intersegmentarias son hiperintensas y nítidas, mientras que en la RM estándar permanecen difusas o invisibles.

En la búsqueda por optimizar estas delimitaciones en RM, la arquitectura de red juega un rol crucial. Investigaciones como la de Raab et al. [2] han contrastado variantes de nnU-Net (estándar y con codificador residual) frente a arquitecturas basadas en transformadores de visión (Swin UNETR), demostrando que nnU-Net superó de manera consistente a las alternativas. Además, los transformadores sufrieron una mayor caída de rendimiento al enfrentar imágenes provenientes de distintos escáneres. A pesar de estos avances técnicos y de los niveles de precisión global alcanzados (resumidos en la Tabla 1), persiste una deficiencia crítica para su adopción clínica, que es la falta de cuantificación del error predictivo. Aunque estudios en neuroimagen [16] usan medidas de incertidumbre para detectar anomalías anatómicas, patologías o errores de etiquetado, este análisis es escaso en la segmentación hepática. En esa línea, Mehrtash et al. [31] mostraron que la entropía media sobre la región segmentada se correlaciona con el Dice y detecta segmentaciones deficientes sin referencia, principio que sustenta la correlación de Spearman de la sección 4.5.2. La diferencia está en el objeto,

porque allí se trata de estructuras de borde visible y aquí de fronteras que la imagen no muestra, expresadas en volumen por segmento. Hasta donde alcanza la revisión realizada, no existen estudios previos que evalúen o expresen la incertidumbre algorítmica en los ocho segmentos anatómicos de Couinaud y la traduzcan a métricas de volumen útiles en la práctica médica.

Tabla 1. Desempeño reportado en segmentación hepática sobre RM

Trabajo	Modalidad y datos	Tarea	Dice
CHAOS [18]	T1-DUAL, sujetos sanos	Hígado	0,954
Zbinden et al. 2022 [1]	T1 vibe Dixon sin contraste	Hígado	0,936
Raab et al. 2025 [2]	T1 vibe multifase con contraste	Hígado	0,970
Zbinden et al. 2023 [21]	T1 vibe Dixon sin contraste, 170/30	Segmentos I–VIII	0,770
Han et al. 2022 [22]	Fase portal, 367/157/158	Segmentos	0,920
Miao et al. 2024 [23]	CE-MRI fase portal, 78/10/25	Segmentos	0,853
Nicoli et al. 2025 [24]	RM heterogénea, validación cruzada	Segmentos	0,847

2.3 Trabajos relacionados y delimitación del aporte

El antecedente metodológico directo es [16], que evalúa la incertidumbre en el etiquetado de arterias intracraneales mediante aumentación en tiempo de prueba con inversión guiada por coordenadas y compara tres variantes de U-Net bajo protocolo unificado. Sin embargo, las divergencias con la presente investigación son sustanciales porque el objeto de estudio es un órgano sólido y no una red tubular, lo que altera radicalmente la naturaleza de las fronteras anatómicas y la pertinencia de las métricas topológicas; la entrada al modelo es una imagen representada en una matriz de intensidades radiológicas puras, en lugar de una segmentación binaria previa; y, finalmente, la resolución de los vóxeles manejados son marcadamente anisotrópicos[3].

Frente a este escenario, la investigación plantea una triple contribución:

- Primero, la construcción de una base de datos de RM parcelada con los ocho segmentos de Couinaud, elaborada bajo un protocolo documentado y revisado por un médico radiólogo especialista.
- Segundo, el diseño de una comparación controlada de tres arquitecturas de segmentación empleando un único código de evaluación sobre un conjunto de prueba común lo que permite aislar empíricamente el rendimiento intrínseco de cada red del impacto derivado del efecto de preprocesamiento.
- Tercero, la cuantificación sistemática de la incertidumbre a escala intersegmentaria y su respectiva proyección matemática a unidades volumétricas de uso clínico, abordando un área en la que, hasta donde alcanza la revisión realizada, no se registran antecedentes formales.

3 Desarrollo

3.1 Base de datos

Se construyó una base sobre estudios de RM hepática procedentes del conjunto público Duke Liver Dataset [26]. Las máscaras correspondientes a los ocho segmentos fueron delineadas manualmente empleando el software 3D Slicer, mediante la subdivisión de la máscara global del hígado proporcionada en el conjunto original. Las etiquetas de los segmentos constituyen una partición espacial exacta de una delimitación hepática

[3] Anisotrópico: vóxel cuyas dimensiones difieren según la dirección; típicamente, la distancia entre cortes es mayor que la resolución dentro del corte.

previamente validada. Las dimensiones de las matrices variaron entre 256 × 256 × 72 y 512 × 512 × 80 vóxeles, con espaciados típicos de 0,74–1,64 mm en el plano y 2,5–4,0 mm entre cortes. Esta configuración originó una elevada anisotropía (mediana de 4,65 y máximo de 8,97), observándose que 117 de los 197 volúmenes superaron una relación anisotrópica de 3. Todos los volúmenes se encuentran en orientación axial (transversal), con codificación LPS (cuyos ejes espaciales positivos apuntan hacia las direcciones anatómicas Izquierda, Posterior y Superior del paciente) siendo un esquema de codificación nativo en el entorno de 3D Slicer que garantiza la coherencia espacial y anatómica de las entradas al modelo. De los 310 volúmenes disponibles se descartaron 113 (36%) en el control visual previo a la anotación: 105 por calidad de imagen insuficiente y 8 por anomalías que impedían distinguir las venas que delimitan los planos intersegmentarios. De los 197 restantes, 131 son adquisiciones sin contraste (precontraste con supresión grasa, en fase y en fase opuesta) y 65 con contraste, en su mayoría de fase portal; el conjunto de prueba mantiene esa proporción. La composición final de la base se resume en la Tabla 2.

La división se hizo por paciente y no por volumen, ya que cada paciente aporta en promedio tres volúmenes y separarlos habría permitido que la red fuera evaluada sobre un hígado que ya conocía, elevando las métricas sin reflejar un mejor desempeño. Los 73 pacientes se repartieron en un 70% para entrenamiento, un 15% para validación y un 15% para prueba, con una semilla fija que se guardó para que el reparto pudiera repetirse y fuera el mismo en las tres arquitecturas y en las implementaciones II y III. Los casos de prueba se apartaron desde el comienzo y no se usaron para elegir modelos ni ajustar parámetros.

Tabla 2. Composición de la base de datos

Concepto	Valor
Volúmenes disponibles	310
Volúmenes descartados en el control visual	113
Volúmenes segmentados	197
Pacientes	73
Volúmenes por paciente (mediana)	3
Emparejamiento imagen–máscara	197 de 197
Casos con los ocho segmentos	182
Casos con anotación parcial	15
Entrenamiento	136 volúmenes, 51 pacientes
Validación	31 volúmenes, 11 pacientes
Prueba	30 volúmenes, 11 pacientes

3.2 Construcción de la base utilizando protocolo de anotación manual

La construcción del conjunto etiquetado constituyó el primer objetivo específico y demandó el mayor esfuerzo en horas-humanas. Fue ejecutada por dos alumnos en práctica, quienes documentaron el proceso en texto y video, bajo supervisión del radiólogo vinculado al proyecto, quien validó el protocolo de anotación y revisó parte de los volúmenes, sin registro del número revisado. Ambos trabajaron sobre los mismos estudios y la concordancia entre sus anotaciones no se midió. Sobre los archivos DICOM importados a 3D Slicer, cada anotador examinó los cortes en sentido craneocaudal hasta identificar las tres venas hepáticas y trazar una línea desde el origen de cada una, marcando como referencia la terminación de la vena cava. El volumen se seccionó a lo largo de esas líneas, y un corte transversal en el nivel marcado separó los segmentos superiores de los inferiores. El lóbulo caudado (segmento I) fue la única excepción, ya que se delimitó manualmente al final, por su tamaño reducido y su variabilidad morfológica. El procedimiento fue iterativo. En una primera versión la tercera línea seguía el trayecto de la vena, pero la revisión radiológica mostró que el segmento IV invadía los territorios del II y del III, de modo que en la segunda se extendió hasta el pliegue donde se originan las venas restantes y la separación se refinó manualmente, estableciéndose esa versión como referencia.

Tres consideraciones metodológicas condicionan la interpretación de los resultados: la subdivisión se realizó mediante planos geométricos y no siguiendo estrictamente los territorios portales; el uso de un único corte transversal para todo el órgano es una simplificación anatómica; y la numeración que genera el software difiere de la clasificación clásica de Couinaud, lo que requirió un mapeo de correspondencia (sección 3.3).

3.3 Desarrollo de la solución: implementaciones sucesivas

El desarrollo siguió una secuencia de etapas, en que cada una surgía de los problemas detectados en la anterior, comenzando con una etapa exploratoria que fijó el diseño experimental (implementación I), seguida de dos implementaciones definitivas para la delimitación de los ocho segmentos de Couinaud: una a resolución original (implementación II), y la otra sobre una caja isotrópica de 64^3 (implementación III), finalizando con la reconstrucción de la red de hígado completo.

El proceso se inició con una etapa exploratoria cuyo propósito no fue maximizar el rendimiento predictivo, sino comprender la naturaleza de los datos y establecer el diseño experimental. Para ello, se construyó un flujo en PyTorch (implementación I) con decisiones de diseño adoptadas manualmente: los volúmenes se alinearon anatómicamente, se remuestrearon a una dimensión fija de $64 \times 64 \times 64$ vóxeles, y se normalizó su intensidad recortando los valores atípicos (por debajo del percentil 0,5 y por encima del 99,5). Esta primera versión se entrenó con el optimizador AdamW [27] con planificación coseno y una función de pérdida que combinó entropía cruzada y coeficiente de Dice. Esa fase sentó las bases metodológicas para las implementaciones posteriores, definiendo aspectos clave como la comparación de arquitecturas bajo un protocolo unificado, la selección de métricas espaciales y el análisis de sensibilidad e incertidumbre. En particular, la decisión de realizar la partición del conjunto de datos a nivel de paciente (y no por imagen) fue fundamental y se conservó durante todo el desarrollo. Dado que un paciente puede aportar múltiples volúmenes con distintos contrastes, separarlos de forma aislada habría provocado contaminación entre los conjuntos de entrenamiento y prueba, lo que mediría la memorización de la red en lugar de su capacidad real de generalización.

A partir de esta etapa inicial surgieron dos hallazgos que motivaron la reestructuración del procedimiento. En primer lugar, forzar todos los volúmenes a un mismo número de vóxeles los deforma de manera desigual. Al contar con espaciados anisótropicos (del orden de $0{,}82 \times 0{,}82 \times 4$ mm), cada estudio se comprimía en distinta proporción, alterando la morfología aparente del hígado y la red debía aprender esas deformaciones además de la anatomía. Este hallazgo determinó que todas las implementaciones posteriores definieran la región de interés por su extensión física en milímetros y no por un conteo de vóxeles. En segundo lugar, se detectó que la numeración de algunas de las etiquetas generadas en 3D Slicer no coincidía con la segmentación de Couinaud, discrepancia que se corrigió inspeccionando la posición anatómica y validando el volumen relativo de cada región.

Para dar solución a estos problemas, se desarrolló una segunda iteración, que corresponde a la implementación II, trabajando con el marco nnU-Net v2 [11]. La decisión se fundamenta en tres aspectos. Primero, el marco deriva automáticamente el plan de preprocesamiento a partir de las características del conjunto de datos, lo que corrige directamente la deformación detectada en la etapa anterior; Segundo, evalúa las predicciones en el espacio de resolución original, condición necesaria para expresar las métricas en milímetros y centímetros cúbicos; y tercero, constituye el estándar de referencia en segmentación médica, lo que permite contrastar los resultados con la literatura. En esta fase, se fijó la partición definitiva por paciente y estratificada según la presencia de cada segmento. Los 73 pacientes se repartieron en 51 para entrenamiento, 11 para validación y 11 para prueba, con 136, 31 y 30 volúmenes respectivamente, verificando que ningún paciente apareciese en más de un subconjunto.

Las tres arquitecturas se identifican con la nomenclatura de sus respectivos marcos. La U-Net 3D corresponde a la configuración que nnU-Net deriva por defecto del conjunto de datos. ResEnc M es la variante con codificador residual en su tamaño intermedio, dentro de una familia de tres configuraciones (S, M y L) que difieren en el número de canales y en los requerimientos de memoria. MedNeXt-B es la variante de tamaño

base de esa arquitectura, dentro de una familia de cuatro tamaños (S, B, M y L), y el sufijo k3 indica un núcleo de convolución de 3 × 3 × 3.

Bajo este esquema, se entrenaron las tres arquitecturas empleando en cada una de ellas 250 épocas (un cuarto de las 1.000 que nnU-Net aplica por omisión) en un único entrenamiento por arquitectura, y la misma partición sobre una GPU NVIDIA L4 en Google Colab, con guardado y reanudación automática ante cortes de sesión (Tabla 3). Cabe mencionar que el modelo MedNeXt-B requirió de un tratamiento particular, ya que al depender de nnU-Net v1, incompatible con la versión 2, se ejecutó en un entorno virtual aislado y se configuró con un plan de espaciado anisotrópico para evitar interpolaciones excesivas entre cortes. Finalmente, todas las redes se evaluaron sobre su espacio anatómico original, con un mismo código de evaluación, y sus predicciones se estandarizaron mediante un postproceso que extrae únicamente la componente conexa mayor, eliminando así posibles ruidos en la segmentación.

A partir del mismo conjunto etiquetado se construyó una tercera versión, que corresponde a la implementación III, diseñada específicamente para el estudio de incertidumbre. En este caso, cada estudio se enmarcó en una caja cúbica de 288 mm centrada en el hígado y se remuestreó a una dimensión de 64^3 con vóxeles isotrópicos de 4,5 mm. El centro se calculó sobre la componente conexa mayor de la anotación y no sobre su totalidad, de modo que eventuales fragmentos aislados de etiqueta no desplazaran la caja. La imagen se suavizó antes de reducir la resolución y la etiqueta se trasladó clase por clase, conservando en cada vóxel nuevo la de mayor fracción de volumen, para no perder segmentos pequeños como el I. El conjunto de prueba, la partición y el mapa de etiquetas son los mismos de la implementación II, de manera que ambas resultan comparables volumen por volumen.

Esta configuración ofreció importantes ventajas prácticas, ya que con vóxeles isotrópicos una rotación de 10° significa lo mismo en los tres ejes, lo que permite comparar la sensibilidad entre ejes sin que el resultado quede dominado por la anisotropía. Además, el límite superior geométrico se mantiene sobre 0,98 incluso a 30°, de modo que la degradación observada es atribuible al modelo y no al remuestreo. Por último, el costo computacional del análisis (veintiocho predicciones por volumen, por red) resulta abordable en unos minutos en lugar de varias horas.

Tabla 3. Configuración de las tres arquitecturas en las implementaciones II y III

Red	Espaciado (mm)	Parche (vóxeles)	Entrenamiento
II: U-Net 3D	4,0 × 0,82 × 0,82	32 × 224 × 256	7,1 h
II: ResEnc M	4,0 × 0,82 × 0,82	32 × 192 × 224	9,3 h
II: MedNeXt-B (k=3)	4,0 × 0,82 × 0,82	32 × 224 × 256	19,8 h
III: U-Net 3D	4,5 isotrópico	64 × 64 × 64	≈ 2,8 h
III: ResEnc M	4,5 isotrópico	64 × 64 × 64	≈ 2,8 h
III: MedNeXt-B (k=3)	4,5 isotrópico	64 × 64 × 64	≈ 3 h

Las seis: 250 épocas, con la misma partición por paciente y checkpoint de mejor validación

Finalmente, el desarrollo se complementó con la reconstrucción de un modelo auxiliar de segmentación de hígado completo heredado de un trabajo previo y no reproducible en su estado original ya que dependía de rutas fijas, no declaraba su partición y sus pesos no podían recargarse por diferencias de nomenclatura entre versiones. Se trata de una U-Net 3D que trabaja sobre volúmenes de 64 × 64 × 64 y distingue únicamente hígado de fondo, entrenada en dos etapas. La primera fue un autoencoder que comparte codificador y embotellamiento, sobre 1.753 volúmenes sin anotación del Duke Liver Dataset, y la segunda, la segmentación propiamente tal con los 207 volúmenes anotados de entrenamiento, durante 100 épocas con optimizador Adam y pérdida de Dice. El modelo alcanzó un coeficiente de Dice de 0,803 sobre sus 35 volúmenes de prueba. Cabe aclarar que este modelo no forma parte de la evaluación comparativa, ya que su tarea es binaria y tanto su partición como su cohorte difieren de las implementaciones II y III, encargadas de subdividir el órgano en ocho segmentos.

Adicionalmente, los archivos generados por este modelo no conservan la resolución espacial en milímetros, motivo por el cual los volúmenes se expresan en vóxeles y no en centímetros cúbicos. La inclusión de esta red alternativa tiene como único propósito someterla a la misma evaluación de incertidumbre sobre el hígado completo, ejecutada sobre sus propios volúmenes de prueba (apartados de su entrenamiento) y no sobre los 30 de las implementaciones II y III, por lo que la comparación es cualitativa y no pareada (sección 4.5.2).

3.4 Estimación de la incertidumbre

El procedimiento sigue a [16], adaptado a imágenes de intensidad (Figura 3). Para cada volumen de prueba y cada red se obtienen K + 1 predicciones, la original y K predicciones sobre versiones transformadas de la imagen, las cuales se revierten al espacio original antes de agregarlas.

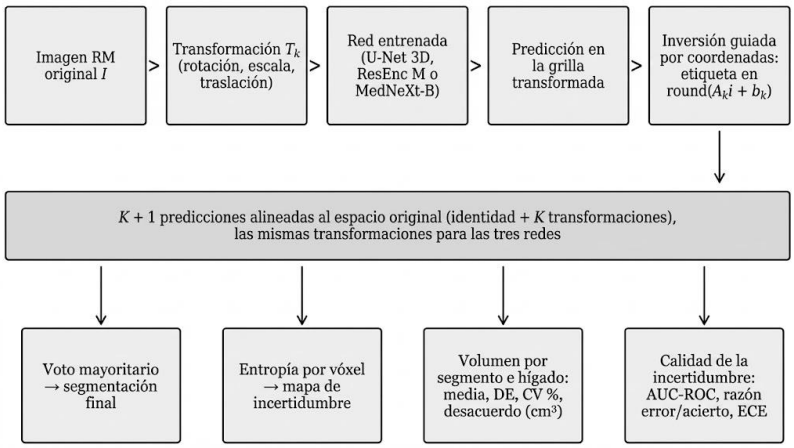


Figura 3. Procedimiento de estimación de incertidumbre por aumentación en tiempo de prueba con inversión guiada por coordenadas[4].

Transformaciones en espacio físico. Cada transformación combina una rotación R, un escalado S y una traslación t. Para respetar el espaciado anatómico, estas se definen en milímetros y se aplican sobre los índices de vóxel i (fórmula 1):

$$i' = A \cdot i + b, A = D^{-1} S\ R\ D \tag{1}$$

donde D es la matriz diagonal de espaciados y b el desplazamiento en índices de vóxel, que reúne la traslación t y el término que sitúa el centro del volumen como punto de rotación. La imagen se proyecta a la grilla transformada mediante interpolación trilineal, ampliando la grilla lo necesario para evitar que la rotación recorte el órgano.

Inversión de transformaciones guiada por coordenadas. La predicción de la red, generada sobre la grilla transformada, se devuelve a su espacio original sin interpolar las etiquetas, ya que interpolar valores discretos introduce errores precisamente en las fronteras, que es donde se mide la incertidumbre. Para cada vóxel original, se asigna la etiqueta predicha en la posición redondeada de su imagen (fórmula 2):

$$P_k(i) = \hat{Y}_k\big(\mathrm{round}(A_k i + b_k)\big) \tag{2}$$

[4] El diagrama ilustra el flujo de trabajo desde la imagen de RM original hasta las métricas finales. Se aplican K transformaciones geométricas a la imagen, cada versión transformada se procesa en la red entrenada y las predicciones resultantes se revierten al espacio original, generando un conjunto de K+1 predicciones alineadas, de las cuales se derivan la segmentación final, los mapas de entropía y las métricas de validación volumétrica y de incertidumbre.

Este enfoque equivale al método guiado por coordenadas descrito en [16]. A diferencia de dicho trabajo, en esta investigación la red produce la etiqueta directamente sobre la grilla transformada, por lo que este redondeo es el único paso discreto. Para aislar el error atribuible exclusivamente a la manipulación de la imagen, se calculó un límite superior geométrico. La referencia manual se proyectó a la grilla transformada y se devolvió al espacio original. El Dice resultante representa el rendimiento máximo alcanzable tras esa transformación, suponiendo una red perfecta.

Conjuntos de transformaciones. Para estimar la **incertidumbre** se emplearon K = 8 transformaciones aleatorias por volumen (además de la predicción original): rotaciones de hasta ±10° en el plano axial y ±5° fuera de él, escalados entre 0,9 y 1,1 y traslaciones de hasta ±5 vóxeles en el plano y ±1 corte. Estas mismas transformaciones se aplicaron a las tres redes evaluadas para garantizar comparaciones pareadas. Por otro lado, para evaluar la **sensibilidad** se aplicaron rotaciones deterministas de 5°, 10°, 15°, 20° y 30° sobre cada eje por separado (Z cráneo-caudal, Y antero-posterior y X izquierda-derecha) y escalados de 0,8, 0,9, 1,1 y 1,2. En ambos experimentos se desactivó el efecto espejo interno de nnU-Net durante la inferencia para aislar el efecto de las transformaciones propias; por este motivo, el coeficiente de Dice de la predicción original en esta etapa es ligeramente inferior al reportado en la sección 4.3.

Definición de métricas de evaluación. Para cuantificar el desempeño y la confianza del modelo, los resultados se analizaron en tres niveles:

- **Medidas por vóxel.** Con las $K+1$ predicciones alineadas, se obtiene la segmentación final por voto mayoritario. A partir de la fracción de votos $p(c|i)$ de cada clase c, se calcula la entropía normalizada (fórmula 3):

$$H(i) = -\frac{1}{\ln 9}\sum_c p\,(c|i)\cdot \ln p\,(c|i) \tag{3}$$

  Este valor es 0 cuando todas las predicciones coinciden y 1 ante un desacuerdo máximo entre las nueve clases (fondo y ocho segmentos). Para el hígado completo, se emplea la entropía binaria, calculada sobre: $p(hígado|i) = 1 - p(fondo|i)$.

- **Medidas volumétricas por segmento.** Para cada segmento c y para el hígado completo, se calcula el volumen $Vk(c)$ en cada una de las $K+1$ predicciones y se reportan su media, desviación estándar (DE) y coeficiente de variación (fórmula 4):

$$CV(c) = 100\cdot \frac{\mathrm{DE}[V_0(c),\ldots,V_K(c)]}{\mathrm{media}[V_0(c),\ldots,V_K(c)]} \tag{4}$$

  Adicionalmente, se calcula el **volumen de desacuerdo**, compuesto por los vóxeles que al menos una predicción asigna al segmento c y al menos otra no. Estas medidas responden a una pregunta clínica directa: ¿Cuántos centímetros cúbicos de un segmento son inciertos?

- **Calidad de la incertidumbre.** Para evitar que los vóxeles de fondo sobreestimen las métricas, la calidad de la estimación se evalúa solo dentro de la unión de la región predicha y la de referencia. Para ello, se utilizan tres indicadores: el AUC-ROC de detección, la razón entre la entropía media de los vóxeles erróneos y la de los correctos, y el ECE. Para este último se toma como confianza la fracción de votos de la clase ganadora, repartida en diez intervalos de igual ancho. Este ECE mide por tanto la calibración del consenso entre las nueve predicciones de la TTA, y no la del softmax de la red, que se discutió en la sección 1.2. El error de calibración esperado se calcula entonces (fórmula 5):

$$ECE = \sum_b \frac{n_b}{N}\cdot |\mathrm{exactitud}(b) - \mathrm{confianza}(b)| \tag{5}$$

  Complementariamente, se registra qué fracción de la incertidumbre y del error se localiza en la frontera

interna entre segmentos, en el borde externo del hígado o en su interior. Esta región de evaluación requiere la referencia manual. En cambio, la entropía media por estudio que se emplea para priorizar la revisión (sección 4.5.2) se promedia únicamente sobre la región predicha.

- **Análisis estadístico.** Las comparaciones entre redes se realizaron volumen por volumen mediante la prueba de Wilcoxon pareada [28], aplicando la corrección de Holm [29]. Los intervalos de confianza del 95% se estimaron por *Bootstrap*, remuestreando por paciente, ya que los 30 volúmenes de prueba provienen de solo 11 pacientes y tratarlos como independientes habría estrechado los intervalos de forma artificial. Las diferencias de incertidumbre entre segmentos se contrastaron con la prueba de Friedman (aplicada a los volúmenes que presentaban los ocho segmentos completos), utilizando la W de Kendall como medida del tamaño del efecto. Para evitar definiciones circulares, los volúmenes con anotación parcial se excluyeron del análisis por segmento según la presencia de etiquetas en la referencia manual, y no según el resultado de la métrica. Wilcoxon, Friedman y Spearman suponen independencia, que los volúmenes de prueba no cumplen estrictamente, lo que se declara como limitación.

- **Implementación y puntos de control.** El procedimiento se consolidó en un único programa que se conecta a los modelos previamente entrenados (sin necesidad de reentrenamiento) y guarda la información en almacenamiento persistente por cada red y por volumen por separado, de modo que un corte en la sesión no implique pérdida de progreso. Finalmente, este mismo programa se aplicó, con una sola red y salida binaria, a la U-Net 3D propia de hígado completo (sección 4.5.2).

4 Resultado y análisis

4.1 Efecto de la preparación de los datos

Las tres implementaciones comparten conjunto de prueba, partición por paciente y familias de arquitecturas, y se diferencian en la preparación de los datos. La implementación I fija un número constante de vóxeles sin considerar el tamaño físico que representan. La implementación III fija una caja de 288 mm centrada en el hígado y la reduce a ese mismo número de vóxeles, de 4,5 mm cada uno. Por su parte, la implementación II conserva la resolución original del plano, con vóxeles de 4,0 × 0,82 × 0,82 mm. La Tabla 4 las dispone en ese orden, que no es cronológico del desarrollo sino el que permite aislar un factor a la vez. Entre la primera y la segunda columna solo cambia la definición geométrica porque ambas trabajan con 64^3 vóxeles, y entre la segunda y la tercera solo cambia la resolución.

Tabla 4. Dice medio por segmento sobre los 30 volúmenes de prueba según la preparación de los datos (media ± DE)

Arquitectura	I: número fijo de vóxeles	III: caja de 64^3 a 4,5 mm	II: resolución original	I → III	III → II
U-Net 3D	0,5556 ± 0,1131	0,7470 ± 0,1084	0,7739 ± 0,0824	+0,1914	+0,0269
ResEnc M	0,6775 ± 0,1040	0,7417 ± 0,1074	0,7782 ± 0,0845	+0,0642	+0,0365
MedNeXt	0,6152 ± 0,1346	0,7440 ± 0,1208	0,7580 ± 0,0972	+0,1288	+0,0140

Pasar de la implementación I a la III mejora el Dice entre 0,064 y 0,191 sin cambiar la cantidad total de vóxeles. El paso a la II multiplica esa cantidad por más de treinta, y aporta un Dice entre 0,014 y 0,036, mientras que el cambio de arquitectura mueve el resultado entre 0,004 y 0,020. La ganancia se explica mejor por no deformar cada estudio que por disponer de más resolución. No obstante, entre la implementación I y III cambia también el marco de trabajo, de modo que la diferencia observada refleja el procedimiento completo y no puede atribuirse únicamente a la definición geométrica. La evaluación que sigue se apoya en las implementaciones II y III.

4.2 Segmentación del hígado completo

Los resultados obtenidos en la segmentación del hígado completo (Tabla 5) revelan que las seis configuraciones superan el umbral de Dice de 0,93. Este nivel de desempeño se encuentra dentro del rango reportado en la literatura especializada para imágenes de resonancia magnética (RM). En particular, se alinea con el valor de 0,936 documentado por Zbinden et al. [1] sobre adquisiciones tipo Dixon sin medio de contraste. Delimitar el hígado, por tanto, no es la parte difícil de esta tarea. Las diferencias entre las dos resoluciones no superan el 0,005 y cambian de orden según la arquitectura, lo que confirma que ninguna arquitectura presenta ventaja sistemática en el contorno externo.

Tabla 5. Segmentación del hígado completo sobre los 30 volúmenes de prueba

Implementación	Arquitectura	Dice	HD95 (mm)	ECE
II	U-Net 3D	0,9357	—	—
II	ResEnc M	0,9373	—	—
II	MedNeXt-B	0,9329	—	—
III	U-Net 3D	0,9362	7,94	0,0937
III	ResEnc M	0,9334	9,06	0,0917
III	MedNeXt-B	0,9381	7,76	0,0927

En la implementación III, el Dice corresponde a la evaluación del conjunto de prueba, mientras que las métricas HD95 y ECE provienen del análisis de incertidumbre (ejecutado con el efecto espejo interno desactivado). Bajo esta condición, los valores de Dice para el hígado son 0,9334 en la U-Net 3D, 0,9296 en ResEnc M, y 0,9352 en MedNeXt-B. El HD95 y el ASSD de la Tabla 6 promedian los segmentos, no el contorno hepático, que no se calculó en la implementación II. El ECE se calculó sobre la fracción de votos derivada del proceso de aumentación

4.3 Segmentación de los segmentos de Couinaud

A resolución original (implementación II), la arquitectura ResEnc M alcanzó el mayor coeficiente de Dice medio con 0,778, seguida estrechamente por la U-Net 3D con 0,774 y, en último lugar, por MedNeXt-B con 0,758 (Tabla 6). La prueba pareada no detectó una diferencia significativa entre ResEnc M y la U-Net 3D (Δ = 0,004; $p = 0,318$), aunque ambas superaron de forma consistente a MedNeXt-B (Δ = 0,020 y 0,016; $p < 0,001$ y $p = 0,042$ tras la corrección de Holm). En la caja isotrópica de 64^3 (implementación III), las tres quedan dentro de un Dice de 0,006 entre sí y ninguna comparación resultó significativa ($p \geq 0,42$ tras aplicar la corrección de Holm). Esto evidencia que, al reducir la resolución espacial, las diferencias entre arquitecturas dejan de ser detectables.

Tabla 6. Segmentos de Couinaud sobre los 30 volúmenes de prueba: resolución original (II) y caja isotrópica de 64^3 (III)

Implement.	Red	Dice medio [IC 95%]	DE	Mediana	HD95 (mm)	ASSD (mm)
II	U-Net 3D	0,774 [0,744; 0,801]	0,082	0,805	13,01	3,93
II	ResEnc M	0,778 [0,747; 0,806]	0,085	0,805	13,48	4,50
II	MedNeXt-B	0,758 [0,722; 0,790]	0,097	0,784	14,53	4,37
III	U-Net 3D	0,747 [0,707; 0,783]	0,108	0,768	14,48	5,07
III	ResEnc M	0,742 [0,703; 0,778]	0,107	0,760	14,89	5,14
III	MedNeXt-B	0,744 [0,699; 0,783]	0,121	0,784	15,35	5,08

Los intervalos de confianza (IC) se estimaron mediante remuestreo (bootstrap) agrupado por paciente. Las métricas de distancia superficial (HD95 y ASSD) se reportan promediadas exclusivamente sobre los segmentos anatómicos presentes en la referencia manual de cada volumen evaluado

La validación interna de cada marco de trabajo evidenció un sesgo metodológico recurrente que el diseño experimental de este estudio logró mitigar. Según sus propios números, MedNeXt-B parecía la mejor red por casi tres centésimas de Dice (0,794 frente a 0,767 de ResEnc M); sin embargo, medida sobre el conjunto de

prueba con el mismo código de evaluación que las otras dos redes, quedó última dentro de la implementación II. Entre ambas cifras cambian a la vez el conjunto evaluado, el código que calcula las métricas y el criterio de selección del punto de control, de modo que la diferencia no es atribuible a una sola causa. Medir las tres redes con un único código sobre un conjunto de datos estandarizado es lo que sostiene la validez de estas comparaciones. El nivel de complejidad lo marca el segmento anatómico subyacente y el patrón se repite en ambas resoluciones (Tabla 7). Por ejemplo, el segmento I, cuyo volumen promedio de 48,4 cm³ es aproximadamente cinco veces menor que el de sus vecinos, registra sistemáticamente los peores desempeños en las seis configuraciones, lo cual es esperable porque en una región pequeña cualquier desviación de frontera pesa proporcionalmente más.

Tabla 7. Dice por segmento de Couinaud en prueba y volumen medio de referencia

Segmento	n	II U-Net 3D	II ResEnc M	II MedNeXt-B	III U-Net 3D	III ResEnc M	III MedNeXt-B	Volumen (cm³)
S1	30	0,716	0,726	0,724	0,701	0,689	0,707	48,4
S2	30	0,784	0,771	0,765	0,721	0,740	0,739	252,3
S3	28	0,732	0,726	0,706	0,682	0,689	0,639	151,0
S4	30	0,742	0,771	0,753	0,733	0,727	0,719	233,6
S5	30	0,804	0,808	0,794	0,793	0,779	0,787	293,4
S6	30	0,849	0,847	0,836	0,812	0,789	0,819	263,5
S7	29	0,769	0,775	0,724	0,747	0,739	0,757	242,8
S8	29	0,794	0,801	0,759	0,790	0,784	0,783	288,1

El volumen corresponde a la referencia manual medida en la caja de 64³. Respecto de la resolución original difiere menos de un 2% en todos los segmentos

Sin embargo, el segmento III rompe este patrón volumétrico. Pese a su tamaño intermedio (151,0 cm³), presenta un bajo rendimiento sistemático y no está delimitado en la referencia de dos volúmenes del mismo paciente (n=28); en uno de ellos faltan también el VII y el VIII, los más craneales, lo que apunta a un campo de visión que no cubre todo el hígado y no a una ausencia anatómica. Ambos están entre los 15 casos con anotación parcial de la Tabla 2. Para ilustrarlo, la Figura 4 presenta tres casos de prueba: una segmentación típica, un error de asignación espacial entre territorios adyacentes (generado por la U-Net 3D) y un caso de alta dificultad anatómica (donde el Dice indicado corresponde al volumen completo).

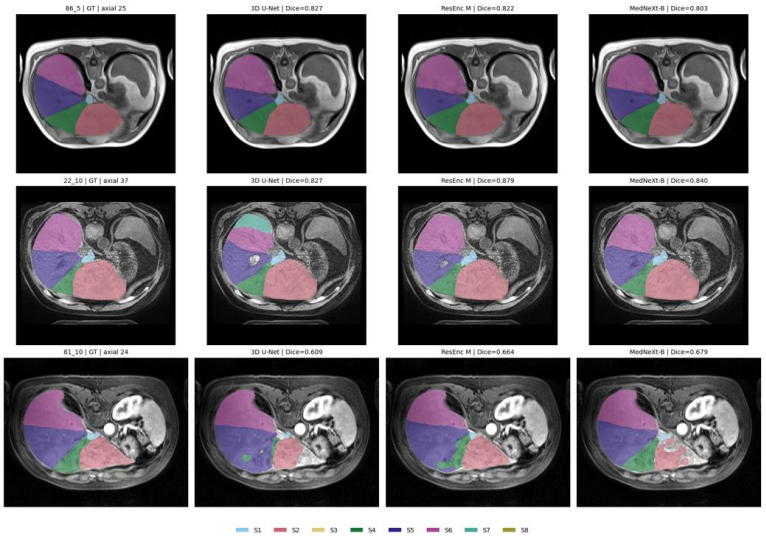


Figura 4. Casos representativos de prueba: típico (arriba), error de asignación entre territorios de la U-Net 3D (centro) y caso difícil (abajo). El Dice indicado corresponde al volumen completo

La brecha de 16 puntos entre el hígado completo (0,937) y sus segmentos (0,778) es el costo de repartir el hígado en ocho territorios sin fronteras radiológicas visibles. Al no haber bordes que detectar, la red infiere dónde suelen ir los planos a partir de referencias vasculares, y ahí es donde las arquitecturas dejan de diferenciarse. Asimismo, el error de contorno (entre 3,9 y 5,1 mm) equivale a un corte radiológico, señal de que operan en el límite de resolución del dato. El Dice de la segmentación interna queda apenas dos puntos por debajo de 0,80, plenamente alineado con el referente más cercano en RM sin contraste (Dice de 0,770 con 170 volúmenes) [21], condición que comparten dos tercios de los volúmenes de esta base. Los valores superiores en la literatura corresponden a condiciones distintas: Dice de 0,85 a 0,92 en RM con contraste o con conjuntos de datos mucho mayores [22], [23], y sobre 0,88 en tomografía computarizada (TC), donde las venas que delimitan los segmentos se visualizan con alta nitidez [24], por lo que no son comparables con esta investigación.

4.4 Sensibilidad a transformaciones geométricas

El análisis de sensibilidad geométrica se ejecutó sobre la implementación III, evaluando tres ejes espaciales, cinco ángulos de rotación y cuatro factores de escala sobre los 30 volúmenes de prueba en las tres arquitecturas, utilizando el límite superior geométrico como límite superior de referencia (Figura 5). El resultado separa con nitidez ambas tareas de segmentación. Por un lado, la segmentación del hígado completo apenas pierde entre 0,03 y 0,09 de Dice incluso a 30°; por otro la subdivisión interna experimentó un deterioro severo, ya que, al rotar 30° sobre el eje anteroposterior (eje Y), el desempeño de la U-Net 3D descendió drásticamente de 0,736 a 0,500 (una pérdida relativa del 32%), y MedNeXt-B cayó de 0,742 a 0,519. Como el límite superior geométrico se mantuvo entre 0,978 y 1,000 en todas las condiciones evaluadas, se concluye que esta pérdida de precisión proviene de las limitaciones propias de los modelos frente a los cambios de posición, y no de un error matemático al procesar la imagen. Ambas arquitecturas conservaron la aumentación por omisión de nnU-Net, que para los parches isotrópicos de 64^3 aplica rotaciones de hasta ±30° en cada uno de los tres ejes, escalados entre 0,7 y 1,4 y espejo en los tres. Los ángulos evaluados quedan por tanto dentro del rango visto durante el entrenamiento, de modo que la degradación no obedece a extrapolación sino a una generalización incompleta dentro del propio rango de aumentación.

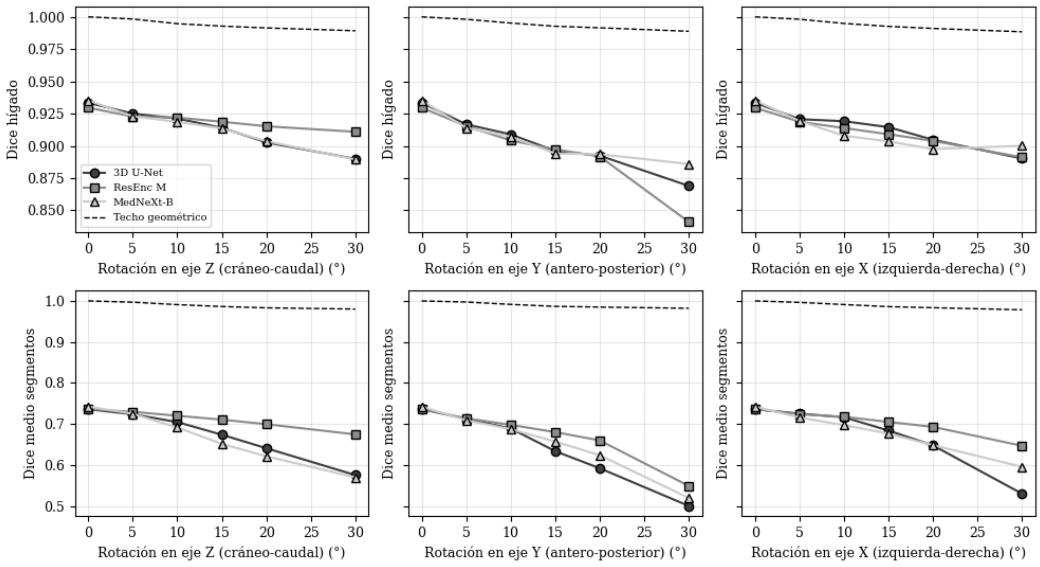


Figura 5. Dice del hígado completo (arriba) y medio de los segmentos (abajo) tras rotar sobre cada eje y revertir la rotación, para las tres redes de la implementación III; la línea discontinua es el límite superior geométrico

Por ejes, la rotación anteroposterior resultó la más perjudicial por razones anatómicas: inclina el plano coronal y desplaza las venas hepáticas que delimitan los planos intersegmentarios. En contraste, el giro craneocaudal

(eje Z) fue el menos perjudicado al conservar la relación espacial en el plano axial (ResEnc M mantuvo un Dice de 0,674 a 30°, frente al 0,500 de la U-Net 3D). Los cambios de tamaño generaron un impacto mucho menor, ya que reducir la imagen al 80% restó apenas entre 0,020 y 0,032 de Dice, y ampliar al 120% movió el resultado entre +0,003 y −0,016. El efecto se aprecia mejor en volumen que en Dice. Sobre el conjunto de rotaciones evaluadas, de 5° a 30° en los tres ejes, el volumen del hígado completo varió entre un 4% y un 5%, el del segmento I entre un 21% y un 42% y el del III entre un 24% y un 47%. La dependencia con el ángulo es directa porque una inclinación de solo 5° ya desplaza en promedio un 15% el volumen de ambos segmentos, frente al 2,6% del hígado, y a 30° los segmentos superaron el 50%. Es decir, un mismo estudio radiológico, adquirido con el paciente levemente inclinado, puede producir predicciones volumétricas segmentarias que difieren en más de un tercio. Esta pronunciada variabilidad fija un límite clínico restrictivo sobre el grado de confianza que puede otorgarse actualmente a una volumetría segmentaria automática.

4.5 Evaluación de incertidumbre

La evaluación de la incertidumbre constituye el objeto central de este trabajo y se ejecutó sobre la implementación III (ocho segmentos e hígado completo). Las tres arquitecturas se sometieron a transformaciones idénticas, volumen por volumen, lo que asegura comparaciones estrictamente pareadas. De forma independiente se replicó el análisis sobre el modelo auxiliar de hígado completo (ajeno a nnU-Net), para verificar que los hallazgos no dependen del marco empleado (sección 4.5.2). La implementación II, a resolución original, se excluyó por su excesivo costo computacional (horas frente a minutos), y queda planteada como trabajo futuro.

4.5.1 Calidad de la incertidumbre por arquitectura

Las tres redes producen estimaciones de incertidumbre prácticamente indistinguibles (Tabla 8). El AUC-ROC de los segmentos varía entre 0,715 y 0,725, y a nivel de hígado entre 0,805 y 0,816, con intervalos de confianza que se solapan por completo. La incertidumbre detecta el error bastante mejor en el hígado que en los segmentos, y la razón está en la naturaleza de cada uno. El error del contorno externo es una banda estrecha donde las predicciones discrepan entre sí, mientras que el error intersegmentario desplaza un plano completo, dejando extensas regiones donde todas las predicciones coinciden y se equivocan a la vez.

Tabla 8. Implementación III: calidad de la incertidumbre por arquitectura (30 volúmenes, 11 pacientes)

Red	DICE sin TTA segm.	DICE con TTA segm.	AUC-ROC segm.	AUC-ROC hígado	Razón segm.	Razón hígado	ECE segm.	ECE hígado
3D U-Net	0,736	0,722	0,716 [0,687; 0,742]	0,805 [0,784; 0,825]	3,08	9,04	0,205	0,094
ResEnc M	0,736	0,729	0,715 [0,691; 0,738]	0,813 [0,791; 0,830]	3,45	9,53	0,212	0,092
MedNeXt-B	0,742	0,711	0,725 [0,696; 0,754]	0,816 [0,786; 0,841]	2,98	8,33	0,203	0,093

La razón corresponde al cociente entre la entropía media de los vóxeles erróneos y la de los correctos. IC 95% (bootstrap). Sin diferencias significativas entre redes (Wilcoxon con corrección de Holm; $p \geq 0{,}21$ en AUC-ROC y $p \geq 0{,}34$ en ECE)

La razón de entropía entre vóxeles erróneos y correctos es de 2,98 a 3,45 en los segmentos, y de 8,33 a 9,53 en el hígado. El ECE de los segmentos (entre 0,203 y 0,212) es alto, lo que revela una sobreconfianza sistemática, pues la fracción de votos sobrestima la exactitud, por lo que una coincidencia total entre predicciones no garantiza una respuesta correcta. A nivel de hígado, en cambio, el ECE es más de dos veces mejor (≈0,093). En cuanto a su distribución, más del 90% de la incertidumbre se concentra en las fronteras, con un 50% a 53% en las divisiones internas y cerca de un 40% en el borde externo. En consecuencia, la cobertura del error (proporción de vóxeles mal clasificados con discrepancia predictiva) se sitúa entre el 62% y el 69%. Aproximadamente un tercio de los fallos pasa inadvertido.

El voto mayoritario por TTA degradó sistemáticamente el desempeño en las tres redes. El Dice de los segmentos bajó de 0,736 a 0,722 en la U-Net 3D; de 0,736 a 0,729 en ResEnc M; y de 0,742 a 0,711 en MedNeXt-B. Asimismo, el HD95 empeoró entre 0,5 y 3,1 mm. A diferencia de lo observado en estructuras tubulares [16], aquí la TTA empeora las métricas. Este resultado es consistente con la sensibilidad geométrica (sección 4.4) porque si el modelo se degrada ante pequeñas rotaciones, promediar predicciones rotadas mezcla segmentaciones de peor calidad con la original. Por tanto, en este problema, la TTA debe entenderse como un instrumento para medir la incertidumbre, no para mejorar la segmentación.

4.5.2 Incertidumbre por segmento y en volumen

El volumen del hígado completo se estima con un sesgo aceptable del 2% al 4% (Tabla 9) y su distribución junto a la de los segmentos se presenta en la Figura 6. En los segmentos aparecen errores sistemáticos. El segmento I se subestima en torno a un 27% en las tres redes, mientras que el IV se sobreestima en media entre un 28% y un 50%. Ambos sesgos son coherentes con el protocolo de anotación, ya que la frontera del lóbulo caudado (segmento I) se delinea a mano, y el segmento IV limita con el II y el III justamente en el plano que requirió correcciones manuales en la primera versión del protocolo.

Tabla 9. Volúmenes (cm³, media ± DE) entre referencia manual vs voto mayoritario por red (Implementación III)

Segmento	n	Referencia	U-Net 3D	Error %	ResEnc M	Error %	MedNeXt-B	Error %
S1	30	48,4 ± 29,8	36,1 ± 24,6	−29,4	37,2 ± 23,9	−25,3	38,1 ± 27,2	−28,7
S2	30	252,3 ± 99,7	220,4 ± 80,1	−10,3	240,5 ± 77,2	+1,3	229,8 ± 79,9	−3,1
S3	28	151,0 ± 91,0	152,9 ± 81,3	+17,2	149,2 ± 76,4	+9,9	148,3 ± 94,8	+2,7
S4	30	233,6 ± 134,3	263,3 ± 93,2	+39,7	251,7 ± 85,2	+28,1	273,4 ± 108,2	+50,4
S5	30	293,4 ± 98,1	335,6 ± 117,2	+19,1	324,4 ± 107,6	+15,9	341,1 ± 106,7	+21,5
S6	30	263,5 ± 103,4	270,7 ± 97,5	+6,2	268,3 ± 100,7	+6,1	268,6 ± 90,5	+6,7
S7	29	242,8 ± 160,1	198,3 ± 95,1	+6,9	222,6 ± 110,5	+17,4	188,3 ± 109,4	−6,5
S8	29	288,1 ± 127,4	298,3 ± 109,9	+6,9	302,4 ± 107,6	+9,4	285,5 ± 119,6	+0,2
Hígado	30	1745,4 ± 549,9	1763,8 ± 493,6	+2,5	1786,0 ± 503,7	+3,6	1759,3 ± 490,5	+2,3

El error porcentual es la media de los errores relativos por caso, no la razón de las medias; por eso un segmento puede tener medias parecidas y error medio alto. En el S7 la mediana del error es −20,4 %, −15,2 % y −16,8 %: el segmento se subestima y la media positiva proviene de casos aislados de hasta +274 %.

La Figura 7 presenta la distribución de los volúmenes predichos por cada arquitectura, conservando los ejes de la Figura 6 para facilitar la comparación directa con la referencia manual.

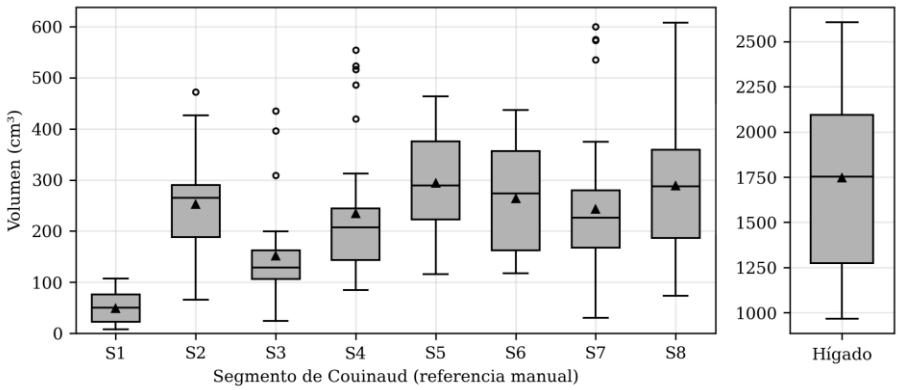


Figura 6. Distribución del volumen de referencia para los segmentos de Couinaud y el hígado completo

Los tres paneles resultan prácticamente superponibles, lo que confirma en términos volumétricos la equivalencia estadística ya observada en el Dice. Adicionalmente, las tres redes replican el patrón de la referencia, con una dispersión reducida en el segmento I y mayor en el V y el VIII. Estiman, además, el volumen del hígado completo en un rango notablemente más estrecho que el de sus segmentos individuales.

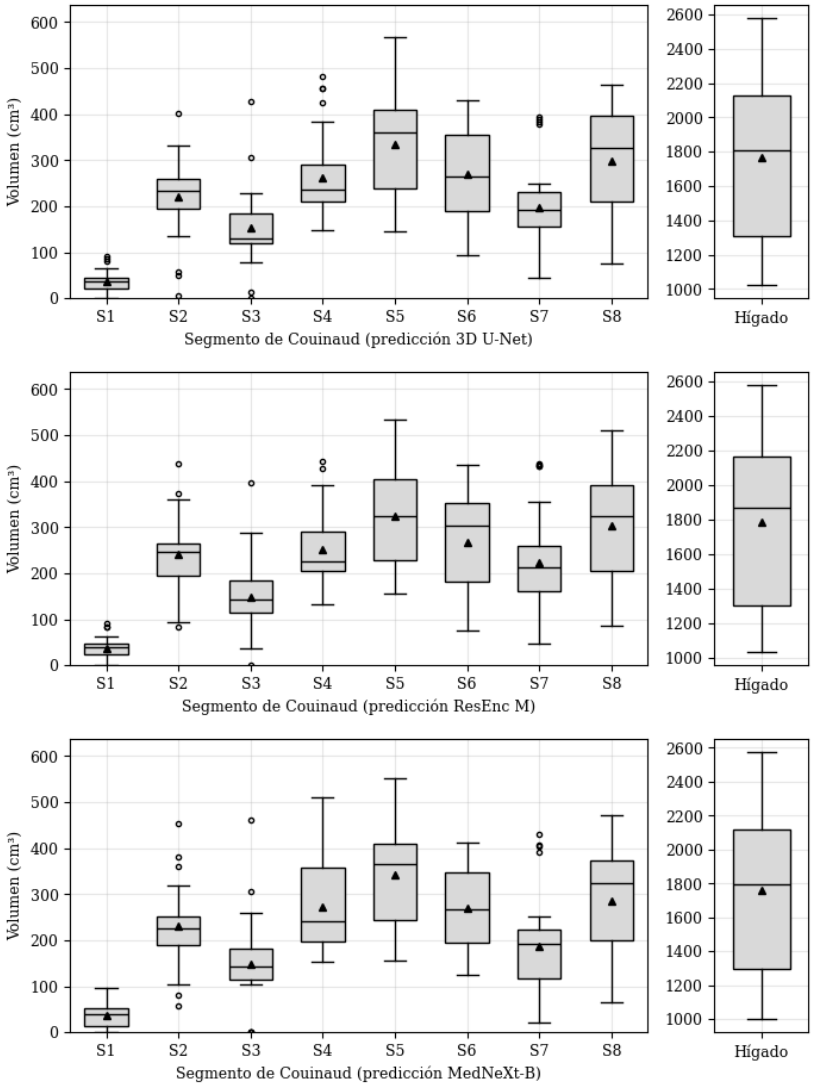


Figura 7. Distribución de los volúmenes segmentarios predichos por las tres arquitecturas evaluadas

La incertidumbre volumétrica ordena los segmentos de manera nítida y estable (Tabla 10 y Figura 8). Mientras que el hígado completo varía apenas un 3% entre predicciones, los segmentos internos fluctúan entre un 5% y un 39%. Los segmentos más inciertos son sistemáticamente el I (25%-37%) y el VII (9%-19%), seguidos del III (15%-39%); en contraste, los más estables son el VIII (6%-9%) y el V (7%-10%). Este ordenamiento depende del tamaño porque tanto el CV como la entropía media correlacionan con el volumen de referencia (ρ entre $-0{,}69$ y $-0{,}93$; $p \leq 0{,}06$) y el segmento I, el más pequeño, es el de menor desacuerdo absoluto. El orden expresa error relativo, no inestabilidad intrínseca.

Los segmentos I y III coinciden además con los de peor Dice; sin embargo, las discrepancias entre ambas métricas resultan informativas. El segmento VII presenta alta incertidumbre, pero un Dice intermedio, reflejando su ubicación en la cúpula hepática (donde suele terminar el campo de visión). El segmento II muestra

el caso opuesto, ya que exhibe un Dice bajo y poca incertidumbre, señal de un sesgo estable en el modelo que la TTA no logra revelar. Este último es el ejemplo concreto del límite del método y explica por qué el AUC-ROC de los segmentos se detiene en torno a 0,72.

Tabla 10. Incertidumbre volumétrica por segmento entre las K + 1 predicciones aumentadas (media sobre casos)

Segmento	DE (cm³) U-Net 3D	DE (cm³) ResEnc M	DE (cm³) MedNeXt-B	CV % U-Net 3D	CV % ResEnc M	CV % MedNeXt-B
S1	3,95	3,84	4,06	30,6 (9,1)	25,2 (8,1)	37,0 (9,1)
S2	17,72	18,67	23,64	12,9 (6,2)	9,3 (5,8)	11,2 (7,0)
S3	16,59	13,05	17,46	23,4 (7,8)	15,4 (7,1)	38,6 (8,2)
S4	26,16	19,16	33,98	9,6 (8,4)	7,3 (6,6)	11,0 (9,2)
S5	24,56	24,20	37,49	7,7 (5,9)	7,2 (4,6)	10,3 (5,9)
S6	21,73	23,49	30,30	9,1 (7,2)	9,2 (4,6)	11,8 (7,2)
S7	21,79	16,91	26,62	12,6 (10,2)	8,8 (7,0)	19,0 (13,8)
S8	21,10	15,40	21,96	7,7 (4,9)	5,5 (4,0)	9,2 (6,3)
Hígado	46,65	59,87	48,21	3,0 (1,2)	3,7 (1,2)	3,1 (1,1)

DE y CV: desviación estándar y coeficiente de variación volumétrica (9 predicciones). El CV se expresa como media (mediana) sobre casos; la distribución es asimétrica (un caso alcanza 300%) y el ordenamiento entre segmentos se conserva con ambos. Volumen de desacuerdo: 33,8-35,1 cm³ (S1), 115,3-128,2 cm³ (S3) y 562,7-620,8 cm³ (hígado completo)

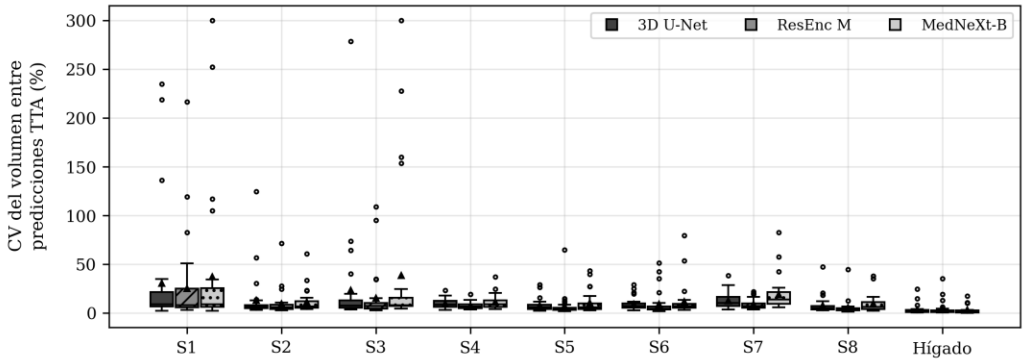


Figura 8. Coeficiente de variación del volumen de cada segmento entre las predicciones aumentadas, por red

La prueba de Friedman (Tabla 11) muestra que estas diferencias no son atribuibles al azar ($p < 10^{-6}$), y sitúa al segmento I como el más dudoso en las tres redes.

Tabla 11. Ordenamiento de los segmentos por incertidumbre (prueba de Friedman sobre los 28 volúmenes completos)

Red	χ^2	p	W de Kendall	Orden, de más a menos incierto
U-Net 3D	54,67	$1{,}7 \times 10^{-9}$	0,279	S1 > S7 > S3 > S2 > S4 > S8 > S6 > S5
ResEnc M	66,75	$6{,}7 \times 10^{-12}$	0,341	S1 > S7 > S4 > S3 > S2 > S6 > S8 > S5
MedNeXt-B	55,87	$1{,}0 \times 10^{-9}$	0,285	S1 > S7 > S4 > S2 > S3 > S8 > S6 > S5

Contraste sobre entropía media. Los criterios no coinciden: el S3 en MedNeXt-B es el más incierto por CV (tabla 10) y quinto aquí; el CV promedia y Friedman ordena por rangos

La correspondencia espacial entre las zonas de mayor entropía y los errores reales de segmentación se aprecia en los mapas de la Figura 9. Más allá de esa lectura visual, el resultado con mayor proyección clínica es que la entropía promediada sobre todo el volumen, es decir, un único número por estudio, anticipa la calidad de su segmentación. La correlación de Spearman entre el Dice y la entropía media sobre la región predicha, calculada sin usar la referencia manual, oscila entre −0,61 y −0,74 ($p < 0,001$) para las tres redes; los 30 volúmenes provienen de 11 pacientes, por lo que no son observaciones independientes. Esto constituye la principal aplicación práctica del método, ya que permite ordenar automáticamente las listas de trabajo y dirigir la revisión médica hacia los estudios problemáticos sin requerir anotaciones manuales previas. Como verificación adicional, el análisis se replicó en el modelo auxiliar de hígado completo (ajeno a nnU-Net). Su comportamiento coincide con el de las redes principales, con el 96,5% de la incertidumbre en los bordes y una variación volumétrica equivalente del 4,7%, lo que indica que los hallazgos no dependen del marco empleado. Este modelo aportó una observación pertinente para la validación médica: su Dice de 0,803, inferior al rango de 0,92 a 0,97 del estado del arte, convive con una sobreestimación sistemática del 25,5 % en el volumen hepático. El Dice no informa por sí solo sobre el sesgo volumétrico.

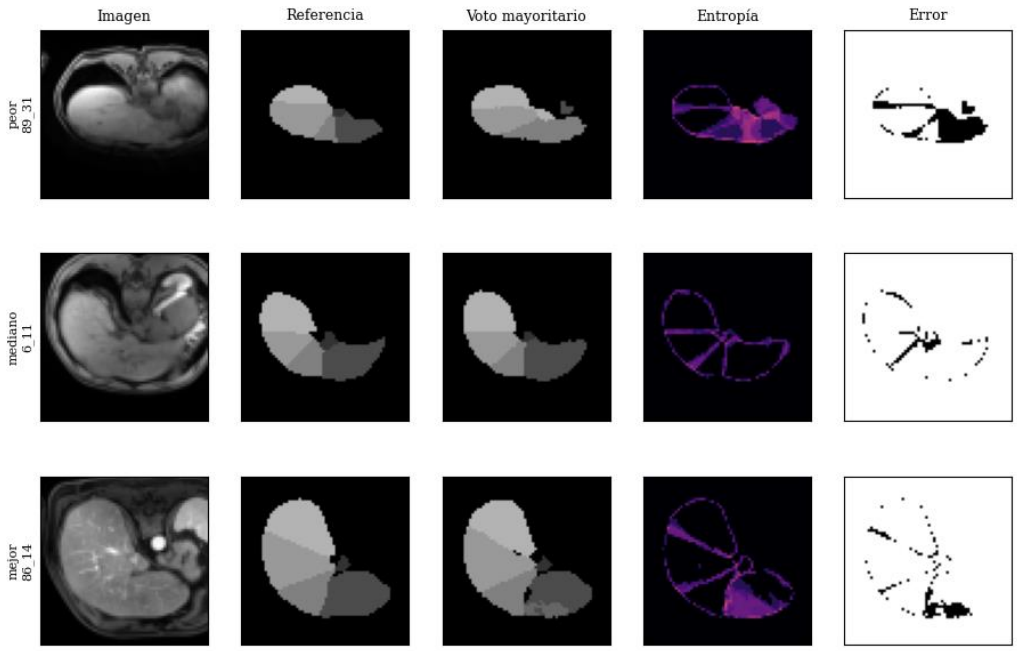


Figura 9. Mapas de incertidumbre de MedNeXt-B, la red de mayor Dice sin TTA en la implementación III, sobre tres volúmenes de prueba: imagen, referencia, voto mayoritario, entropía y error

4.6 Validación del procedimiento de inversión

La inversión de las transformaciones guiada por coordenadas reasignó apenas un 0,38% de los vóxeles de la referencia en el conjunto de transformaciones de aumentación, frente al 1,07% producido por el remuestreo estándar mediante el vecino más cercano, es decir, cerca de un tercio de ese error. El límite superior geométrico se mantuvo en 0,993 ± 0,007 para dicho conjunto, y en 0,985 ± 0,017 al incluir los escalados espaciales. Este control permite atribuir la degradación del rendimiento observada en la sección 4.4 al modelo predictivo y no a la manipulación geométrica de los datos. Sin él, una fracción de lo que se reportaría como duda de la red correspondería, en realidad, a ruido introducido por el método de medición.

4.7 Identificación de hallazgos

A partir de la evaluación de los modelos, se consolidan los siguientes hallazgos:

- **Sobre la preparación de los datos:** La coherencia geométrica pesa más que la cantidad de vóxeles. Mantener el tamaño físico de la caja al reducir la resolución apenas degrada el Dice de los segmentos entre 0,014 a 0,036, y no lo afecta en el hígado completo, mientras que deformar los estudios fijando el número de vóxeles empeora el rendimiento de forma sustancial.

- **Sobre la arquitectura:** Su elección incide menos que la estrategia de preparación de los datos. Las diferencias de Dice entre redes alcanzan como máximo 0,020 a resolución original y 0,006 en la caja de 64^3, y su capacidad para detectar el error mediante la incertidumbre resulta estadísticamente indistinguible.

- **Sobre la sensibilidad geométrica**: El contorno del hígado completo es robusto ante rotaciones, pero su subdivisión no. Rotaciones de 30° cuestan hasta un tercio del Dice segmentario y alteran el volumen de los segmentos I y III en más de un 33%. Esos giros no modifican la anatomía, pero exceden la variabilidad habitual de posicionamiento: ya en el rango de ±10° empleado en la TTA, la subdivisión pierde unos 0,047 de Dice sobre el eje antero-posterior frente a unos 0,026 del contorno, cerca del doble (Figura 5).

- **Sobre la incertidumbre:** Detecta eficazmente errores de contorno externo (AUC-ROC de 0,81) pero de forma moderada en el reparto entre segmentos (AUC-ROC de 0,72), ya que el error intersegmentario es un desplazamiento coherente repetido por todas las predicciones. Aun así, cubre hasta el 69% de los errores totales y revela que las redes operan con exceso de confianza (ECE $\approx$ 0,21).

- **Sobre la utilidad práctica:** La correlación de Spearman entre la entropía media sobre la región predicha de cada volumen y su Dice (ρ de $-0{,}61$ a $-0{,}74$) permite priorizar la revisión automatizada de volúmenes dudosos sin requerir anotaciones manuales, identificando consistentemente a los segmentos I y VII como focos de corrección.

- **Sobre la evaluación:** Las métricas internas de cada marco de trabajo no son comparables entre sí y el Dice por sí solo resulta insuficiente. El modelo auxiliar de hígado completo lo ilustra: con un Dice de 0,803, bajo el estado del arte, sobreestimó el volumen real del órgano en un 25,5%.

5 Conclusiones

Este trabajo midió cuánta confianza admite una segmentación hepática automática en resonancia magnética, tanto del hígado completo como de sus ocho segmentos de Couinaud. Dos resultados enmarcan esa medición. El primero es que la preparación de los datos incide más que la elección de la arquitectura. El Dice de las tres redes difiere como máximo en 0,020 entre sí a resolución original y de 0,006 en la caja de 64^3, mientras que fijar el número de vóxeles en lugar de su tamaño físico degrada el resultado de forma sustantiva. El segundo es que, con un Dice de 0,937 en el hígado completo y de 0,778 en los segmentos, el problema dejó de ser detectar bordes y pasó a ser inferir planos que no están en la imagen.

Las tres hipótesis sobre incertidumbre convergen en que la calidad de la estimación no depende de la arquitectura, sino del tipo de error anatómico que se intenta detectar.

- **H1 (Refutada):** Las tres redes cumplen la condición de partida con diferencias de Dice mínimas, pero sus valores de AUC-ROC difieren entre 0,0008 y 0,0101, muy por debajo de los 0,05 exigidos. No existe una arquitectura que sepa "dudar mejor" que las demás.

- **H2 (No evaluable):** Las tres redes presentan valores de ECE indistinguibles, de modo que no hubo una red mejor calibrada con la que poner a prueba la hipótesis. La mayor diferencia de AUC-ROC fue de 0,010, por debajo del umbral de 0,03. Que H1 quede refutada y H2 resulte no evaluable descarta la arquitectura como variable explicativa y traslada la pregunta al tipo de error que se intenta detectar, que es lo que examina H3.

- **H3 (Cumplida parcialmente):** En los segmentos se cumple en la U-Net 3D (razón 3,08; AUC-ROC 0,716) y en ResEnc M (3,45; 0,715), mientras que MedNeXt-B queda apenas por debajo del umbral de la razón (2,98; 0,725). En el hígado completo no se cumple: la razón alcanza entre 8,3 y 9,5 y el AUC-ROC entre 0,805 y 0,816, por encima del umbral de 0,75. Ese contraste entre ambos niveles no estaba anticipado por la hipótesis y constituye el hallazgo principal del trabajo. Por su parte, el modelo auxiliar presenta una razón de entropía de 7,50, intermedia entre la de los segmentos y la de las redes principales sobre el hígado, pero un AUC-ROC de detección de 0,710, inferior incluso al de los segmentos ya que su comportamiento no es intermedio en ambas métricas. El error de contorno es de naturaleza local, y ante la duda de incluir o no un vóxel del borde, las predicciones difieren, desacuerdo que la entropía registra. El error intersegmentario, en cambio, es un desplazamiento coherente de un plano divisorio completo que todas las predicciones aumentadas reproducen por igual. Al existir un error carente de desacuerdo predictivo, el instrumento no logra detectarlo. Esta dinámica explica lógicamente por qué más del 90% de la incertidumbre se concentra en las fronteras, por qué la cobertura global del error se estanca entre el 62% y el 69%, y por qué un ECE segmentario de 0,21 revela una sobreconfianza sistemática de las redes.

La hipótesis general se cumple. La entropía media ordena los estudios por calidad sin anotación previa (ρ de $-0{,}61$ a $-0{,}74$), los mapas localizan las regiones menos confiables e identifican sistemáticamente los segmentos I y VII como focos de corrección, y la confianza se expresa en centímetros cúbicos por segmento (variaciones de hasta 47% en los segmentos frente a 4-5% del hígado completo ante rotaciones). El alcance del método es asimétrico porque valida el contorno hepático, pero no detecta el plano intersegmentario desplazado. El Dice obtenido en el hígado y en sus segmentos se sitúa en el rango publicado para RM sin contraste y las variantes con codificador residual no muestran ventajas significativas. A diferencia de otros trabajos en estructuras tubulares, aquí el método de voto mayoritario degrada levemente el desempeño, porque las fronteras internas del hígado no son observables en la imagen. Las principales limitaciones son las 250 épocas por red, un cuarto de las 1.000 por omisión de nnU-Net, y un único entrenamiento por arquitectura, que mezcla diferencias arquitectónicas con ruido de entrenamiento; un conjunto de prueba de 30 volúmenes de 11 pacientes de un solo centro, que no satisface el supuesto de independencia; y la ausencia de un estudio de concordancia entre anotadores que fije el Dice máximo alcanzable. Además, el análisis se ejecutó a resolución reducida y mide inestabilidad ante transformaciones, no incertidumbre epistémica; se mantuvo el espejo izquierda-derecha en el entrenamiento, que penaliza a los segmentos lateralizados; y no se midió la línea base del softmax de la red, por lo que la ventaja de la TTA sobre ella se apoya en [6] y no en datos propios. En conjunto, la incertidumbre por TTA resulta útil para validar contornos e identificar automáticamente los estudios que requieren revisión. El trabajo futuro debería cuantificar la concordancia entre anotadores, incorporar la delimitación guiada por el árbol vascular contra el error intersegmentario, replicar el análisis a resolución original con rotaciones tridimensionales, contrastar la TTA con la entropía del softmax y con el conjunto de las tres redes, y evaluar si esta priorización reduce los tiempos de revisión del radiólogo.

Agradecimientos. A los profesores Julio Sotelo P. y Ricardo Ñanculef A., por la orientación metodológica sostenida a lo largo de todo el proceso, y al radiólogo Dr. Rodrigo Herrera R., del Hospital Clínico de la Universidad Católica por la validación del protocolo de anotación y sus aportes a la investigación. A los alumnos en práctica, Fernanda Araya Z. y Daniel Quispe M. que ejecutaron el trabajo de segmentación manual, sin el cual esta investigación no habría sido posible. A mi amada Natalia Contreras H., cuyo amor incondicional, empatía y constante empuje fueron mi mayor fortaleza para sostener y llevar a término esta ardua investigación.

6 Referencias

[1] Zbinden, L., Catucci, D., Suter, Y., Berzigotti, A., Ebner, L., Christe, A., Obmann, V.C., Sznitman, R., Huber, A.T., "Convolutional neural network for automated segmentation of the liver and its vessels on non-contrast T1 vibe Dixon acquisitions," in Sci. Rep. Vol. 12, 2022, art. 22059, doi: 10.1038/s41598-022-26328-2.

[2] Raab, F., Strotzer, Q., Stroszczynski, C., Fellner, C., Einspieler, I., Haimerl, M., Lang, E.W., "Automatic segmentation of liver structures in multi-phase MRI using variants of nnU-Net and Swin UNETR," in Sci. Rep. Vol. 15, 2025, art. 25740, doi: 10.1038/s41598-025-07084-5.

[3] Couinaud, C., "Le foie. Études anatomiques et chirurgicales". Masson, Paris, 1957.

[4] Abdel-Misih, S.R.Z., Bloomston, M., "Liver anatomy," in Surg. Clin. North Am. Vol. 90, 2010, pp. 643–653, doi: 10.1016/j.suc.2010.04.017.

[5] Instituto de Salud Pública de Chile, "Estadísticas de trasplantes y lista de espera," ISP Chile, Nov. 2024. [Online]. Disponible en: https://www.ispch.gob.cl/wp-content/uploads/2024/12/11.-Trasplantes-y-Lista-Espera-por-mes-Noviembre-2024-final.pdf. [Accedido: 15-Feb-2025].

[6] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., "On calibration of modern neural networks." in Proc. Int. Conf. Machine Learning (ICML). PMLR, Vol. 70, 2017, pp. 1321–1330, arXiv:1706.04599.

[7] Ronneberger, O., Fischer, P., Brox, T., "U-Net. Convolutional networks for biomedical image segmentation." in Proc. MICCAI. Lecture Notes in Computer Science, Vol. 9351, Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[8] Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., "3D U-Net. Learning dense volumetric segmentation from sparse annotation." in Proc. MICCAI. Lecture Notes in Computer Science, Vol. 9901, Springer, 2016, pp. 424–432, doi: 10.1007/978-3-319-46723-8_49.

[9] Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., "Attention U-Net. Learning where to look for the pancreas." in Proc. Medical Imaging with Deep Learning (MIDL), 2018, arXiv:1804.03999.

[10] Roy, S., Koehler, G., Ulrich, C., Baumgartner, M., Petersen, J., Isensee, F., Jaeger, P.F., Maier-Hein, K.H., "MedNeXt. Transformer-driven scaling of ConvNets for medical image segmentation." in Proc. MICCAI. Lecture Notes in Computer Science, Vol. 14223, Springer, 2023, pp. 405–415, arXiv:2303.09975.

[11] Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., "nnU-Net. A self-configuring method for deep learning-based biomedical image segmentation," in Nat. Methods Vol. 18, 2021, pp. 203–211, doi: 10.1038/s41592-020-01008-z.

[12] Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F., "nnU-Net revisited. A call for rigorous validation in 3D medical image segmentation." in Proc. MICCAI. Lecture Notes in Computer Science, Vol. 15009, Springer, 2024, pp. 488–498, arXiv:2404.09556.

[13] Gal, Y., Ghahramani, Z., "Dropout as a Bayesian approximation. Representing model uncertainty in deep learning." in Proc. Int. Conf. Machine Learning (ICML). PMLR, Vol. 48, 2016, pp. 1050–1059, arXiv:1506.02142.

[14] Lakshminarayanan, B., Pritzel, A., Blundell, C., "Simple and scalable predictive uncertainty estimation using deep ensembles." in Advances in Neural Information Processing Systems (NeurIPS), Vol. 30, 2017, pp. 6402–6413, arXiv:1612.01474.

[15] Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks," in Neurocomputing Vol. 338, 2019, pp. 34–45, doi: 10.1016/j.neucom.2019.01.103.

[16] Bisbal, J., Winter, P., Jofre, S., Ponce, A., Ansari, S.A., Abdalla, R., Markl, M., Welin Odeback, O., Uribe, S., Tejos, C., Sotelo, J., Schnell, S., Marlevi, D., "Uncertainty-aware automated labeling of intracranial arteries using deep learning," in BMC Med. Imaging Vol. 26, 2026, art. 175, doi: 10.1186/s12880-026-02276-5.

[17] Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., Wiesenfarth, M., Kavur, A.E., Sudre, C.H., Baumgartner, M., et al., "Metrics reloaded. Recommendations for image analysis validation," in Nat. Methods Vol. 21, 2024, pp. 195–212, doi: 10.1038/s41592-023-02151-z.

[18] Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.-H., Groza, V., Pham, D.D., Chatterjee, S., Özkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., et al., "CHAOS challenge. Combined (CT-MR) healthy abdominal organ segmentation," in Med. Image Anal. Vol. 69, 2021, art. 101950, doi: 10.1016/j.media.2020.101950.

[19] Oh, N., Kim, J.H., Rhu, J., Jeong, W.K., Choi, G.-S., Kim, J.M., Joh, J.-W., "Automated 3D liver segmentation from hepatobiliary phase MRI for enhanced preoperative planning," in Sci. Rep. Vol. 13, 2023, art. 17605, doi: 10.1038/s41598-023-44736-w.

[20] Hänsch, A., Chlebus, G., Meine, H., Thielke, F., Kock, F., Paulus, T., Abolmaali, N., Schenk, A., "Improving automatic liver tumor segmentation in late-phase MRI using multi-model training and 3D convolutional neural networks," in Sci. Rep. Vol. 12, 2022, art. 12262, doi: 10.1038/s41598-022-16388-9.

[21] Zbinden, L., Catucci, D., Suter, Y., Hulbert, L., Berzigotti, A., Brönnimann, M., Ebner, L., Christe, A., Obmann, V.C., Sznitman, R., Huber, A.T., "Automated liver segmental volume ratio quantification on non-contrast T1-Vibe Dixon liver MRI using deep learning," in Eur. J. Radiol. Vol. 167, 2023, art. 111047, doi: 10.1016/j.ejrad.2023.111047.

[22] Han, X., Wu, X., Wang, S., Xu, L., Xu, H., Zheng, D., Yu, N., Hong, Y., Yu, Z., Yang, D., Yang, Z., "Automated segmentation of liver segment on portal venous phase MR images using a 3D convolutional neural network," in Insights Imaging Vol. 13, 2022, art. 26, doi: 10.1186/s13244-022-01163-1.

[23] Miao, D., Zhao, Y., Ren, X., Dou, M., Yao, Y., Xu, Y., Cui, Y., Liu, A., "A multi-task based deep learning framework with landmark detection for MRI Couinaud segmentation," in IEEE J. Transl. Eng. Health Med. Vol. 12, 2024, pp. 697–710, doi: 10.1109/JTEHM.2024.3491612.

[24] Nicoli, A.P., Bach, M., Wasserthal, J., Indrakanti, A.K., Segeroth, M., Yang, S., Cyriac, J., Boll, D., Wilder-Smith, A.J., "Liver segment and lesion segmentation on CT and MRI. An open-source contribution to TotalSegmentator," in J. Imaging Inform. Med. 2025, doi: 10.1007/s10278-025-01716-y.

[25] Arya, Z., Ridgway, G., Jandor, A., Aljabar, P., "Deep learning-based landmark localisation in the liver for Couinaud segmentation." in Proc. Medical Image Understanding and Analysis (MIUA). Lecture Notes in Computer Science, Vol. 12722, Springer, 2021, pp. 227–237, doi: 10.1007/978-3-030-80432-9_18.

[26] Macdonald, J.A., Zhu, Z., Konkel, B., Mazurowski, M.A., Wiggins, W.F., Bashir, M.R., "Duke Liver Dataset. A publicly available liver MRI dataset with liver segmentation masks and series labels," in Radiol. Artif. Intell. Vol. 5, 2023, art. e220275, doi: 10.1148/ryai.220275.

[27] Loshchilov, I., Hutter, F., "Decoupled weight decay regularization." in Proc. Int. Conf. Learning Representations (ICLR), 2019, arXiv:1711.05101.

[28] Wilcoxon, F., "Individual comparisons by ranking methods," in Biometrics Bull. Vol. 1, 1945, pp. 80–83, doi: 10.2307/3001968.

[29] Holm, S., "A simple sequentially rejective multiple test procedure," in Scand. J. Statist. Vol. 6, 1979, pp. 65–70, JSTOR 4615733.

[30] Cameron, J.L., Sandone, C., "Atlas of gastrointestinal surgery, vol. 1." 2nd edition, BC Decker, Hamilton (ON), 2007, p. 121 (fig. 1) y p. 124 (fig. 2).

[31] Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T., "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," in IEEE Trans. Med. Imaging Vol. 39, 2020, pp. 3868–3878, doi: 10.1109/TMI.2020.3006437.