GRADO EN INGENIERÍA MATEMÁTICA E INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

Detección de tumores de riñón en imágenes TAC mediante técnicas avanzadas de Inteligencia Artificial

Autor: Joaquín Mir Macías
Director: Meritxell Riera i Marín, Deep Learning Engineer en SycaiMedical

Madrid, Junio de 2025

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

"Detección de tumores de riñón en imágenes TAC mediante técnicas avanzadas de Inteligencia Artificial"

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2024/25 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.

AUTOR DEL PROYECTO

Fdo.: Joaquín Mir Macías Fecha: 11/06/2025

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Meritxell Riera i Marín Fecha: 11/06/2025

Agradecimientos

Quiero expresar mi más sincero agradecimiento, en primer lugar, a mi *familia*: mis padres, mi hermana, mis tíos, mis primas y mis abuelos, por su cariño, por acompañarme siempre y por ofrecerme la confianza necesaria para alcanzar cada meta.

Extiendo mi gratitud a mi tutora *Meritxell Riera i Marín* por su dedicación, guía y paciencia a lo largo de todo el proyecto; a *Javier García López* y a todo el equipo de *Sycai Medical* por brindarme la oportunidad de desarrollar este trabajo junto a ellos y por todo el aprendizaje compartido.

Agradezco también a la *Escuela Técnica Superior de Ingeniería – ICAI* los años de formación y crecimiento que me ha proporcionado y, muy especialmente, a *David Contreras*, por su incansable esfuerzo y por haber hecho posible esta titulación.

Detección de tumores de riñón en imágenes TAC mediante técnicas avanzadas de Inteligencia Artificial

Autor: Joaquín Mir Macías **Directora:** Meritxell Riera i Marín
Entidad colaboradora: Sycai Medical – ICAI

Resumen El cáncer renal es una patología con alta prevalencia, cuya detección temprana mediante técnicas avanzadas de inteligencia artificial puede mejorar significativamente el pronóstico de los afectados. Este trabajo presenta una evaluación comparativa de varios modelos basados en *Deep Learning* y *Visión por Ordenador* para la segmentación automática y precisa de tumores renales en imágenes médicas obtenidas mediante tomografía axial computarizada (TAC). Se utilizan los conjuntos de datos públicos KiTS19, KiTS21 y KiTS23, sobre los cuales se analizan diferentes arquitecturas convolucionales, incluyendo variantes clásicas de U-Net (2D y 3D), el método autoconfigurable nnU-Net en sus versiones 2D, 3D y en cascada, una versión U-Net 3D con bloques residuales, el modelo Rel-UNet que incorpora estimación de incertidumbre, y el enfoque AutoML Auto3DSeg de MONAI para segmentación 3D. El rendimiento de los modelos se evalúa utilizando métricas avanzadas tanto de solapamiento como de superficie, incluyendo el coeficiente de Dice, Surface Dice con tolerancia τ, la distancia Hausdorff al percentil 95 (HD95). Los resultados demuestran que las arquitecturas 3D superan significativamente a las 2D en detección de tumores. Este estudio busca identificar las fortalezas y limitaciones de cada enfoque, así como explorar técnicas complementarias como la autoconfiguración de arquitecturas y la estimación de incertidumbre, con el objetivo de avanzar hacia soluciones más fiables y automatizadas que puedan apoyar de forma efectiva el diagnóstico clínico y la toma de decisiones médicas en el contexto del cáncer renal.

Palabras clave: Kits, U-Net, nnU-Net, AutoML, Métricas.

1. Contexto y objetivo

El **carcinoma de células renales** (CCR) supone aproximadamente el 2–3 % de todas las neoplasias del adulto, con más de 430 000 diagnósticos y 180 000 muertes anuales en el mundo. La supervivencia a 5 años supera el 90 % cuando el tumor está confinado al riñón, pero cae drásticamente en estadios avanzados, lo que subraya la relevancia de la *detección temprana*. La **tomografía computarizada contrastada** (TAC) es la técnica de referencia para diagnosticar, estadificar y planificar la resección quirúrgica. Sin embargo, la *segmentación manual* de riñón y lesión en volúmenes 3D es laboriosa (30–45 min por caso), sujeta a variabilidad interobservador y poco escalable en flujos clínicos con cientos de estudios semanales.

La irrupción de la **Inteligencia Artificial** ha demostrado que las redes neuronales convolucionales (CNN) pueden automatizar la segmentación con precisión cercana a expertos, aliviando la carga del radiólogo y habilitando mediciones volumétricas objetivas (volumen tumoral, márgenes de resección, crecimiento longitudinal). En particular, los *Kidney Tumor Segmentation Challenges* (KiTS19, KiTS21 y KiTS23) han proporcionado más de mil TAC con máscaras de referencia, impulsando la investigación comparativa de modelos.

Objetivo. Evaluar experimentalmente las arquitecturas CNN más relevantes y seleccionar la que ofrezca la segmentación *más precisa y fiable* del tumor renal, considerando tanto exactitud volumétrica como robustez e incertidumbre.

2. Modelos evaluados

- **U-Net 2D**: encoder–decoder clásico, siendo un modelo rápido y ligero, pero sin coherencia volumétrica. Obteniendo un ($\text{Dice}_{\text{tumor}}$=0.65).
- **U-Net 3D**: versión volumétrica con conv. 3D; explota la volumetría de las imágenes TAC, con un ($\text{Dice}_{\text{tumor}}$=0.70).
- **U-Net 3D Residual**: introduce bloques residuales tipo ResNet para redes más profundas y estables con un ($\text{Dice}_{\text{tumor}}$=0.72).

- **nnU-Net** (2D, 3D, Cascada): framework auto-configurable que ajusta pre- y posproceso, tamaño de parche y *augmentations*; la cascada coarse-to-fine lidera con $\text{Dice}_{\text{tumor}}$=0.85.
- **Rel-UNet**: parte de una nnU-Net 3D y combina varios *checkpoints* obtenidos durante el propio entrenamiento; este pequeño *ensemble* produce mapas de incertidumbre vóxel a vóxel sin perder precisión (Dice tumoral 0.84).
- **Auto3DSeg**: pipeline AutoML de MONAI que entrena y ensambla SegResNet, DynUNet y Swin-UNETR; máxima precisión global ($\text{Dice}_{\text{tumor}}$=0.87).

3. Metodología experimental

- **Datos**. Se unificaron los volúmenes contrastados de las tres ediciones del *Kidney Tumor Segmentation Challenge*:
  - **KiTS19**: 300 casos con imágenes TAC (210 train, 90 test) con dos etiquetas: riñón y tumor.
  - **KiTS21**: 400 casos con imágenes TAC (300 train, 100 test), ahora con tres clases (riñón, tumor, quiste). Cada ROI fue segmentada por $\times 3$ anotadores y fusionada por consenso, aportando variabilidad inter-observador.
  - **KiTS23**: 599 casos con imágenes TAC (489 públicas, 110 privadas) que combinan fases corticomedular (90 %) y nefrogénica (10 %). Una única delineación revisada por experto con corrección con técnicas avanzadas.

  Todos los estudios comparten matriz 512×512, FOV $\approx$ 350 mm y voxel in-plane 0,5–0,8 mm; el grosor axial varía $1-5$ mm.
- **División estratificada**. Los 1 299 casos consolidados se barajaron por centro y fase de adquisición y se particionaron en 60 % *train* (779), 20 % *val* (260) y 20 % *test* (260).
- **Entrenamiento**.
  - Optimizador **AdamW** ($\beta_1 = 0{,}9,\ \beta_2 = 0{,}999,\ w_d = 10^{-2}$).
  - **Scheduler triangular2**: LR oscila linealmente entre 3×10^{-4} y 1×10^{-6} cada 40 epochs.
  - **Augmentations 3D** ($p = 0{,}85$): rotaciones aleatorias ($\pm15°$), *elastic-deform*, *gamma* shift, *mirror* y recortes aleatorios para balancear clases.
  - *Deep supervision* y *early-stopping* tras 40 epochs sin mejora en la pérdida validada.
- **Métricas de evaluación**. Se calcularon cuatro métricas complementarias sobre el conjunto *test*:
  - **Dice global** $\dfrac{2|P \cap G|}{|P| + |G|}$: promedio riñón+tumor.
  - **Dice tumoral**: Dice aplicado sólo a la etiqueta *tumor*; refleja sensibilidad clínica.
  - **Surface Dice (2 mm)**: fracción de la superficie predicha cuya distancia a la superficie real es ≤ 2 mm; valora la precisión de borde.
  - **HD95**: distancia de Hausdorff al percentil 95; acota el peor error excluyendo outliers.

4. Resultados principales

Modelo	Dice global	Dice tumor	Surface Dice	HD95 [mm]
U-Net 2D	0.83	0.70	0.72	15.0
U-Net 3D	0.88	0.79	0.78	12.0
U-Net 3D Residual	0.89	0.80	0.79	11.0
nnU-Net 2D	0.88	0.78	0.77	10.0
nnU-Net 3D	0.91	0.84	0.83	8.0
nnU-Net Cascada	**0.92**	**0.85**	**0.85**	**7.0**
Rel-UNet	0.91	0.84	0.84	7.0
Auto3DSeg	0.90	0.82	0.82	9.0

Tabla 1: Comparativa de rendimiento medio en el conjunto de test KiTS unificado.

El paso de 2D $\rightarrow$ 3D aumenta 0.09 el Dice tumoral y reduce HD95 3 mm. nnU-Net Cascada logra la mejor superposición y la menor distancia de contorno. Rel-UNet añade mapas de incertidumbre sin penalizar métrica. Auto3DSeg obtiene resultados competitivos con mínima intervención manual.

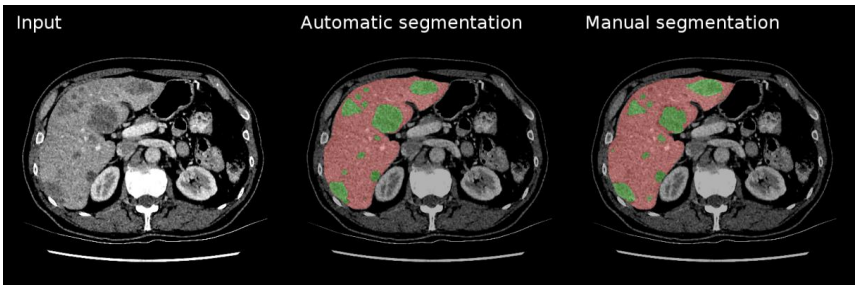

Figura 1: Imagen TAC con segmentación automática y manual

5. Conclusiones

1. Las CNN 3D (U-Net 3D, nnU-Net 3D) superan claramente a los enfoques 2D en segmentación renal.
2. nnU-Net Cascada ofrece la **mayor precisión global** (Dice tumoral 85 %) con un flujo coarse-to-fine.
3. Rel-UNet provee **cuantificación de incertidumbre** útil para control de calidad clínico.
4. Auto3DSeg demuestra que el AutoML puede generar segmentadores sólidos sin ajuste experto.
5. La integración de estas técnicas puede disminuir la carga de segmentación manual y mejorar la planificación quirúrgica en cáncer renal.

Bibliografía resumen [1] F. Isensee *et al.*, "nnU-Net: a self-configuring method for DL-based biomedical image segmentation", *Nat. Methods*, 2021. [2] H. Çiçek *et al.*, "3D U-Net: learning dense volumetric segmentation from sparse annotation", MICCAI 2016. [3] C. Myronenko *et al.*, "Auto3DSeg: Automated 3D segmentation for medical images", MICCAI 2023.

Kidney tumor detection in CT images using advanced Artificial Intelligence techniques

Author: Joaquín Mir Macías **Supervisor:** Meritxell Riera i Marín
Collaborating entity: Sycai Medical – ICAI

Abstract Renal cancer is a highly prevalent disease whose early detection through advanced artificial-intelligence techniques can significantly improve patient prognosis. This work presents a comparative evaluation of several Deep Learning and Computer Vision models for the automatic and precise segmentation of renal tumors in medical images obtained by computed tomography (CT). The public datasets KiTS19, KiTS21 and KiTS23 are used to analyse different convolutional architectures, including classical 2D and 3D U-Net variants, the self-configuring nnU-Net method in its 2D, 3D and cascade versions, a 3D U-Net with residual blocks, the Rel-UNet model that incorporates uncertainty estimation, and MONAI's AutoML Auto3DSeg approach for 3D segmentation. Model performance is assessed using advanced overlap and surface metrics, including the Dice coefficient, Surface Dice with tolerance τ, and the 95th-percentile Hausdorff distance (HD95). The results show that 3D architectures significantly outperform their 2D counterparts in tumor detection. This study seeks to identify the strengths and limitations of each approach, as well as to explore complementary techniques such as architecture self-configuration and uncertainty estimation, with the aim of advancing towards more reliable and automated solutions that can effectively support clinical diagnosis and medical decision-making in the context of renal cancer.

Keywords: KiTS, U-Net, nnU-Net, AutoML, Metrics.

1. Background and objective

Renal cell carcinoma (RCC) accounts for approximately 2–3 % of all adult neoplasms, with more than 430 000 diagnoses and 180 000 deaths annually worldwide. Five-year survival exceeds 90 % when the tumor is confined to the kidney but falls dramatically in advanced stages, underscoring the importance of *early detection*. Contrast-enhanced **computed tomography** (CT) is the reference technique for diagnosis, staging and surgical-resection planning. However, the *manual* segmentation of kidney and lesion in 3D volumes is labour-intensive (30–45 min per case), subject to inter-observer variability and scarcely scalable in clinical workflows that handle hundreds of studies each week.

The advent of **Artificial Intelligence** has shown that convolutional neural networks (CNNs) can automate segmentation with near-expert accuracy, easing the radiologist's workload and enabling objective volumetric measurements (tumor volume, resection margins, longitudinal growth). The *Kidney tumor Segmentation Challenges* (KiTS19, KiTS21 and KiTS23) have provided over one thousand CT scans with reference masks, fuelling comparative model research.

Objective. Experimentally evaluate the most relevant CNN architectures and select the one that offers the *most accurate and reliable* renal-tumor segmentation, considering both volumetric accuracy and robustness/uncertainty.

2. Models evaluated

- **U-Net 2D**: classical encoder–decoder model, fast and lightweight but lacking volumetric coherence. Achieves ($\text{Dice}_{\text{tumor}}$=0.65).
- **U-Net 3D**: volumetric version with 3D convolutions ; exploits CT volumetry, with ($\text{Dice}_{\text{tumor}}$=0.70).
- **U-Net 3D Residual**: introduces ResNet-type residual blocks for deeper, more stable networks with ($\text{Dice}_{\text{tumor}}$=0.72).
- **nnU-Net** (2D, 3D, Cascade): self-configuring framework that tailors pre-/post-processing, patch size and augmentations; the coarse-to-fine cascade leads with $\text{Dice}_{\text{tumor}}$=0.85.
- **Rel-UNet**: builds on a 3D nnU-Net and aggregates several *checkpoints* captured during training; this lightweight ensemble yields voxel-wise uncertainty maps without sacrificing accuracy (tumor Dice 0.84).

- **Auto3DSeg**: MONAI AutoML pipeline that trains and ensembles SegResNet, DynUNet and Swin-UNETR; highest overall accuracy ($\text{Dice}_{\text{tumor}}$=0.87).

3. Experimental methodology

- **Data**. Contrast-enhanced volumes from the three editions of the *Kidney tumor Segmentation Challenge* were unified:
  - **KiTS19**: 300 cases with CT images (210 train, 90 test) with two labels: kidney and tumor.
  - **KiTS21**: 400 cases with CT images (300 train, 100 test), now with three classes (kidney, tumor, cyst). Each ROI was segmented by $\times 3$ annotators and fused by consensus, providing inter-observer variability.
  - **KiTS23**: 599 CT cases (489 public, 110 private) combining corticomedullary (90 %) and nephrogenic (10 %) phases. A single delineation reviewed by an expert with advanced-technique correction.

  All studies share a 512×512 matrix, FOV $\approx$ 350 mm and in-plane voxel $0{,}5-0{,}8$ mm; axial slice thickness varies $1-5$ mm.

- **Stratified split**. The 1 299 consolidated cases were shuffled by centre and acquisition phase and partitioned into 60 % *train* (779), 20 % *val* (260) and 20 % *test* (260).

- **Training**.
  - **AdamW optimiser** ($\beta_1 = 0{,}9,\ \beta_2 = 0{,}999,\ w_d = 10^{-2}$).
  - **Triangular2 scheduler**: LR oscillates linearly between 3×10^{-4} and 1×10^{-6} every 40 epochs.
  - **3D augmentations** ($p = 0{,}85$): random rotations ($\pm15°$), *elastic-deform*, *gamma* shift, *mirror* and random crops to balance classes.
  - *Deep supervision* and *early stopping* after 40 epochs without val-loss improvement.

- **Evaluation metrics**. Four complementary metrics were computed on the *test* set:
  - **Global Dice** $\dfrac{2|P \cap G|}{|P| + |G|}$: average of kidney+tumor.
  - **tumor Dice**: Dice applied only to the *tumor* label; reflects clinical sensitivity.
  - **Surface Dice (2 mm)**: fraction of the predicted surface whose distance to the ground-truth surface is ≤ 2 mm; assesses edge accuracy.
  - **HD95**: 95th-percentile Hausdorff distance; bounds the worst error excluding outliers.

4. Main results

Model	Global Dice	tumor Dice	Surface Dice	HD95 [mm]
U-Net 2D	0.83	0.70	0.72	15.0
U-Net 3D	0.88	0.79	0.78	12.0
U-Net 3D Residual	0.89	0.80	0.79	11.0
nnU-Net 2D	0.88	0.78	0.77	10.0
nnU-Net 3D	0.91	0.84	0.83	8.0
nnU-Net Cascade	**0.92**	**0.85**	**0.85**	**7.0**
Rel-UNet	0.91	0.84	0.84	7.0
Auto3DSeg	0.90	0.82	0.82	9.0

Tabla 2: Average performance comparison on the unified KiTS test set.

Moving from 2D to 3D increases tumor Dice by 0.09 and reduces HD95 by 3 mm. nnU-Net Cascade achieves the best overlap and the smallest contour distance. Rel-UNet adds uncertainty maps without metric penalty. Auto3DSeg attains competitive results with minimal manual intervention.

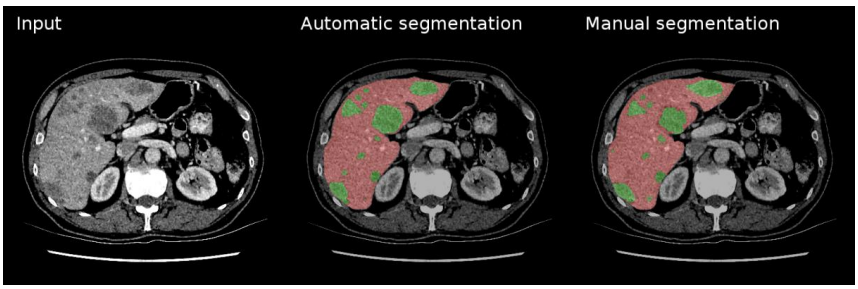

Figura 2: CT image with automatic and manual segmentation

5. Conclusions

1. 3-D CNNs (U-Net 3D, nnU-Net 3D) clearly outperform 2D approaches in renal segmentation.
2. nnU-Net Cascade delivers the **highest overall accuracy** (tumor Dice 85 %) via a coarse-to-fine flow.
3. Rel-UNet provides **uncertainty quantification** useful for clinical quality control.
4. Auto3DSeg shows that AutoML can yield robust segmenters without expert tuning.
5. Integrating these techniques can reduce manual-segmentation workload and improve surgical planning in renal cancer.

Summary bibliography [1] F. Isensee *et al.*, "nnU-Net: a self-configuring method for DL-based biomedical image segmentation", *Nat. Methods*, 2021. [2] H. Çiçek *et al.*, "3D U-Net: learning dense volumetric segmentation from sparse annotation", MICCAI 2016. [3] C. Myronenko *et al.*, "Auto3DSeg: Automated 3D segmentation for medical images", MICCAI 2023.

Índice de la memoria

Índice

1. Introducción

El cáncer renal constituye aproximadamente el **2–3 %** de todas las neoplasias diagnosticadas en adultos, representando un importante problema de salud pública, con más de 430 000 casos nuevos y alrededor de 180 000 muertes al año a nivel mundial [Int24]. El subtipo más frecuente es el *carcinoma de células renales* (CCR), cuya supervivencia a cinco años supera el 90 % cuando el tumor permanece localizado en el riñón, pero disminuye drásticamente si existe diseminación local o metastásica [HSK+19]. La detección temprana y la caracterización precisa de estas lesiones mediante imágenes médicas son, por tanto, esenciales para optimizar las decisiones terapéuticas y mejorar los resultados clínicos.

La tomografía computarizada (TAC) en fase contrastada es la técnica estándar para el diagnóstico y estadificación del CCR. Sin embargo, la segmentación manual de las estructuras anatómicas y tumores en volúmenes 3D es un proceso *laborioso*, *subjetivo* y con una alta variabilidad interobservador [RFB15]. Además, la dificultad para discernir de forma fiable tumores benignos de malignos en imágenes subraya la necesidad de desarrollar métodos automatizados que proporcionen mediciones volumétricas y morfológicas objetivas, útiles tanto en la estratificación del riesgo como en la planificación terapéutica (por ejemplo, elección entre nefrectomía parcial o radical) [IJK+21].

En la última década, las redes neuronales convolucionales (CNNs) han revolucionado la segmentación anatómica en imágenes médicas gracias a su capacidad para aprender patrones espaciales complejos. En particular, la arquitectura U-Net, propuesta inicialmente en 2D y posteriormente extendida a volúmenes 3D, marcó un hito por su precisión y eficiencia en la segmentación médica incluso con conjuntos de datos relativamente reducidos [RFB15]. Modelos posteriores, como la variante auto-configurable nnU-Net, han ampliado estas capacidades automatizando el preprocesamiento, selección de arquitectura y optimización de hiperparámetros, alcanzando un rendimiento competitivo con mínima intervención manual [IJK+21].

Un factor crucial en el desarrollo de estos métodos ha sido la disponibilidad de bases de datos públicas multi-institucionales, destacando especialmente los *Kidney Tumor Segmentation Challenges* (KiTS). Estas bases de datos (KiTS19, KiTS21 y KiTS23) proporcionan cientos de TAC preoperatorias de alta calidad con segmentaciones expertas de tumores renales, quistes y estructuras anatómicas asociadas [HSK+19]. Estas iniciativas han facilitado la creación de benchmarks comunes, permitiendo evaluar de manera sistemática nuevas propuestas y avanzando significativamente en la investigación sobre segmentación renal.

En este trabajo se integra en dicho contexto, explorando y comparando exhaustivamente diversas arquitecturas CNN 3D, incluyendo variantes clásicas y avanzadas de U-Net (2D, 3D, residual), nnU-Net en múltiples configuraciones (2D, 3D y cascada), Rel-UNet con estimación de incertidumbre y AutoML (Auto3DSeg de MONAI). Se implementa una flujo de trabajo reproducible de preprocesado, entrenamiento y evaluación, analizando además el impacto de hiperparámetros clave sobre las métricas más relevantes, como los coeficientes Dice, Jaccard, Surface Dice y la distancia Hausdorff (HD95).

Las principales contribuciones de este trabajo son:

1. Revisar exhaustivamente la **evolución histórica de las CNNs** relevantes en la segmentación médica (Sección 2).

2. Explicación técnica de las **características del entorno de cómputo** utilizado para el entrenamiento y evaluación de los modelos, incluyendo especificaciones de hardware y uso de recursos (Sección 3).

3. Analizar cuantitativamente distintos **datasets TAC públicos** para segmentación renal y sus protocolos de anotación (Sección 4).

4. Evaluar sistemáticamente **múltiples arquitecturas CNN 2D/3D** y métodos AutoML en términos de precisión volumétrica y calidad superficial (Secciones 5–6).

5. Explorar **métricas avanzadas** de evaluación volumétrica y geométrica para segmentación médica (Sección 7).

6. Incorporar **técnicas de explicabilidad e incertidumbre** para mejorar la fiabilidad clínica de los modelos (Sección 8).

Estos resultados destacan el potencial de las técnicas avanzadas de aprendizaje profundo para asistir en la toma de decisiones médicas en cáncer renal, ofreciendo no solo automatización y precisión, sino también métricas robustas para evaluar su desempeño.

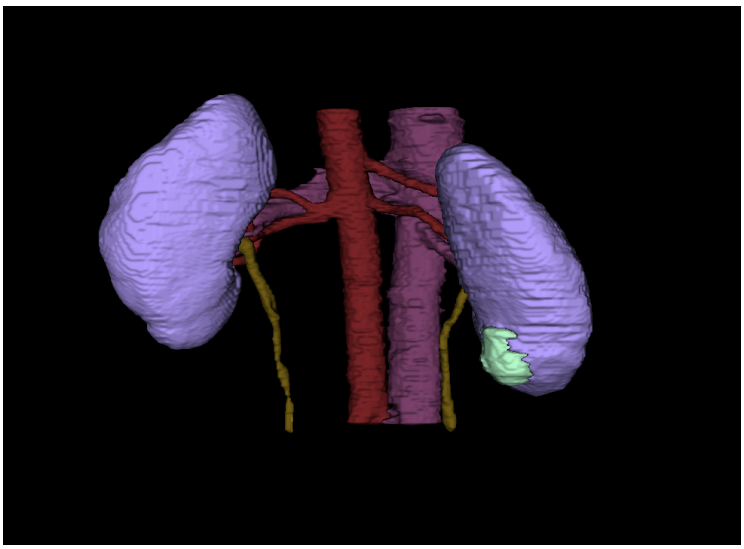

Figura 3: Representación 3D de riñón con tumor.

2. Historia de las Redes Convolucionales

La evolución de las redes neuronales convolucionales (CNNs) ha estado marcada por una sucesión de avances significativos que han ampliado progresivamente su capacidad y eficiencia. Estos desarrollos han tenido un impacto decisivo tanto en la visión por ordenador general como, en los últimos años, en el ámbito específico de la imagen médica. A continuación, se presentan en orden cronológico las arquitecturas y métodos más influyentes:

- **LeNet–5 (1998):** Propuesta por *Yann LeCun* para el reconocimiento de dígitos manuscritos; combinaba *convolución* y *pooling* con 60 k parámetros (unas 430 k conexiones), sentando las bases de las CNN modernas [LBBH98].
- **AlexNet (2012):** Ganó la competición ImageNet 2012 al demostrar el potencial de las CNN profundas (8 capas) con activaciones ReLU, regularización Dropout y entrenamiento en GPU, marcando el despegue del *deep learning* moderno [KSH12].

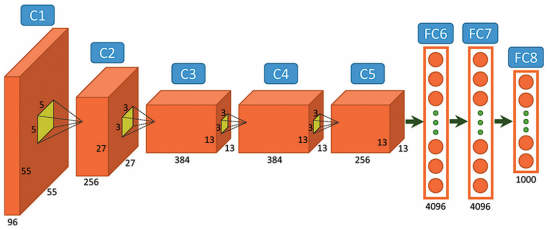

Figura 4: Arquitectura de AlexNet.

- **ZFNet (2013):** Mediante visualización de activaciones ajustó hiperparámetros de AlexNet (filtros $11 \times 11 \rightarrow 7 \times 7$), mejorando la extracción de características y ofreciendo nuevas técnicas de *deconvolutional visualization* [ZF14].
- **Network in Network (NiN, 2013):** Introdujo convoluciones 1×1 y *Global Average Pooling*, incrementando la capacidad no lineal sin disparar los parámetros [LCY14].
- **VGGNet (2014):** Demostró que redes más profundas con filtros 3×3 mejoran la precisión; VGG-16/19 se convirtió en el extractor de características *de referencia* en numerosos trabajos [SZ15].
- **GoogLeNet / Inception (2014):** Incorporó módulos Inception con filtros de varios tamaños en paralelo y proyecciones 1×1, reduciendo parámetros (7 M) y mejorando la eficiencia [SLJ+15].

- **ResNet (2015):** Solucionó el problema del degradado del gradiente mediante *conexiones residuales*, lo que permitió diseñar redes significativamente más profundas y estables [HZRS16].

- **U-Net (2015):** Primer gran hito en imagen biomédica; su arquitectura *encoder–decoder* con *skip connections* permitió segmentar con gran precisión incluso con pocos datos gracias a una agresiva *data augmentation*. Ganó competiciones de microscopía y se convirtió en la base de la mayoría de segmentadores médicos posteriores [RFB15].

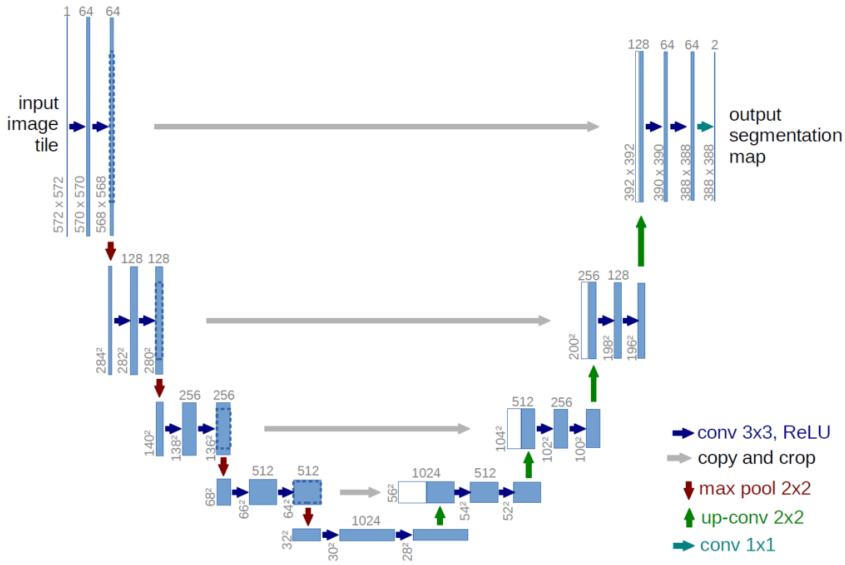

Figura 5: Diagrama de la arquitectura U-Net.

- **3D U-Net (2016):** Çiçek extendieron U-Net al dominio volumétrico, sustituyendo operaciones 2D por 3D para segmentar directamente volúmenes TAC o RMN, clave en radiología [L+16].

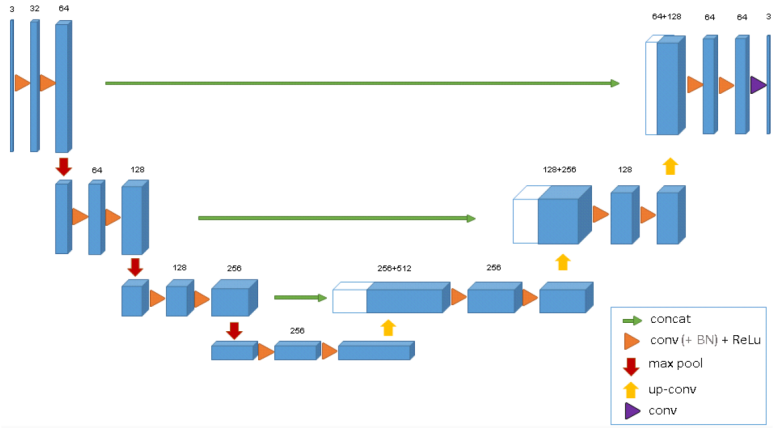

Figura 6: Arquitectura 3D U-Net.

- **V-Net (2016):** Introdujo bloques residuales 3D y la *Dice loss* como función de pérdida, optimizando directamente el coeficiente Dice para manejar el fuerte desbalance entre fondo y lesión en imagen médica [MNA16].

- **FractalNet (2016):** Presentó una estructura jerárquica con regularización *DropPath*, entrenando redes muy profundas sin conexiones residuales [LMS17].

- **DenseNet (2017):** Conectó cada capa con todas las subsiguientes, fomentando la *reutilización de características* y mejorando el flujo de gradientes, reduciendo parámetros [HLVdMW17].

- **nnU-Net (2021):** Fabian Isensee propusieron un marco *auto-configurable* que, a partir de un nuevo dataset, determina de forma automática la resolución, tamaño de parches, hiperparámetros de entrenamiento y *post-processing*. Sin ingeniería manual, ha ganado o quedado entre los cinco primeros en multitud de desafíos de segmentación biomédica [IJK+21].

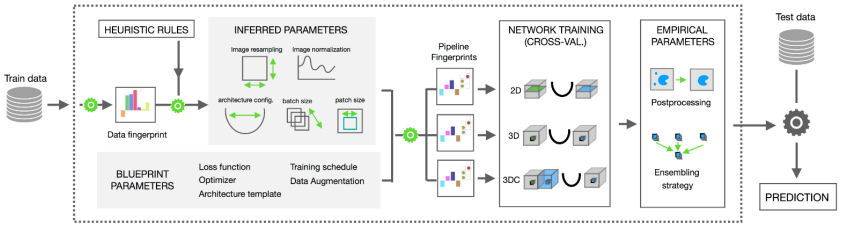

Figura 7: Esquema del framework auto-configurable nnU-Net.

Estas innovaciones – desde la segmentación 2D clásica hasta los métodos volumétricos y la auto-configuración – constituyen la base de los modelos *encoder–decoder* actuales (U-Net 3D, nnU-Net, V-Net, etc.) que dominan la segmentación biomédica. La progresión refleja cómo la comunidad ha pasado de arquitecturas diseñadas ad hoc a marcos automáticos y escalables que se adaptan a distintos conjuntos de datos clínicos, acercando cada vez más el rendimiento de los algoritmos al nivel de los expertos humanos.

3. Entorno de hardware y cómputo

El entrenamiento y la inferencia de los modelos se realizaron en dos estaciones de trabajo con GPUs integradas, lo que permitió reducir significativamente los tiempos de entrenamiento y agilizar la experimentación. La Tabla 3 resume los recursos disponibles en cada equipo. Todas las pruebas se ejecutaron con `Python 3.11.9` sobre `Ubuntu 22.04 LTS`, empleando los controladores `NVIDIA 535.xx` y `CUDA 12.1`.

Tabla 3: Características principales del hardware empleado para la ejecución de los experimentos.

Equipo	**CPU**	**GPU**	**Memoria RAM**	**Almacenamiento**
Lab AI – Aula 602 (ICAI)	Intel Core i7-13700 (13ª Gen)	NVIDIA GeForce RTX 4070 Ti	64 GB	500 GB
Asus ROG Zephyrus G15	AMD Ryzen 9 5900HS	NVIDIA GeForce RTX 3060 Ti	32 GB	600 GB

Notas de uso:

- La mayor parte de los entrenamientos prolongados $\sim$ 10h se llevaron a cabo en el hardware *Lab AI*, aprovechando los 12 GB adicionales de VRAM y el 50 % más de núcleos CUDA de la RTX 4070 Ti. Los modelos más pesados (como nnU-Net 3D o cascada, y Auto3DSeg) fueron ejecutados exclusivamente en este equipo, con tiempos de entrenamiento promedio en torno a las 10 horas por configuración.
- El portátil *Zephyrus G15* se destinó a pruebas rápidas, ajustes de hiperparámetros y validación en movilidad. Además, se realizaron algunas ejecuciones completas de modelos menos pesados (como U-Net 2D y 3D), con duraciones promedio en torno a las $\sim$ 6h.La reproducibilidad entre entornos se garantizó mediante entornos `conda` idénticos.

4. Descripción de los datos e imágenes TAC

Para este estudio se emplean los datasets públicos de la serie *Kidney Tumor Segmentation Challenge* (KiTS), correspondientes a competiciones internacionales de segmentación de tumor renal en TAC. La Tabla 4 resume las principales características de **KiTS19**, **KiTS21** y **KiTS23**: número de casos, clases anatómicas segmentadas y detalles del protocolo de anotación.

Tabla 4: Conjuntos de datos KiTS empleados. Para cada edición se muestran el número de casos (entrenamiento + prueba), las clases segmentadas y un resumen del procedimiento de anotación.

Dataset	Casos	Clases	Protocolo de anotación
KiTS19 [HSK+19]	210 train + 90 test	Riñón, Tumor	Delineación manual única por caso realizada por estudiantes de medicina y revisada por un urólogo oncólogo experto (fase corticomedular).
KiTS21 [H+23a]	300 train + 100 test	Riñón, Tumor, Quiste	Tres delineaciones independientes por región (riñón/tumor/quiste) en plataforma web abierta, con consenso posterior; conjunto de prueba procedente de un centro distinto.
KiTS23 [H+23b]	489 train + 110 test	Riñón, Tumor, Quiste	Una delineación por región realizada por equipo entrenado y revisada por experto; ∼10 % de estudios en fase nefrogénica, el resto en corticomedular.

KiTS19 Incluye 300 estudios TAC de pacientes sometidos a nefrectomía entre 2010 y 2018. Sólo se segmentaron dos clases (riñón y tumor); los quistes renales se integraron dentro de la etiqueta de riñón. Las máscaras se generaron mediante delineado manual en cortes seleccionados con interpolación asistida y posterior revisión experta, garantizando fidelidad clínica.

KiTS21 Añadió la clase *quiste renal*, obteniendo así tres etiquetas por estudio. Cada ROI fue segmentada por tres anotadores distintos a ciegas y posteriormente fusionada, permitiendo medir la variabilidad inter-observador. El conjunto de prueba procede de un hospital distinto al de entrenamiento, lo que introduce un escenario de validación externa.

KiTS23 Reúne 599 estudios (489 públicos + 110 privados para evaluación ciega), integrando todas las anotaciones previas y casos nuevos. A diferencia de KiTS21, se optó por una única delineación revisada por experto tras comprobar que las múltiples delineaciones aportaban beneficio marginal. Introduce mayor heterogeneidad: ∼10 % de los estudios están en fase nefrogénica, aumentando la variabilidad de contraste y la dificultad de la tarea.

Todos los volúmenes KiTS son TAC abdominales con matrices axiales de 512×512 y tamaños de voxel típicamente entre 0,5–0,8 mm in-plane, con espesores de corte de 1–5 mm. Además de las imágenes y máscaras, los datasets proporcionan metadatos clínicos (edad, sexo, tipo de cirugía, etc.). Para este trabajo, todos los conjuntos de datos mencionados anteriormente (KiTS19, KiTS21 y KiTS23) se unificaron en un único dataset consolidado, que posteriormente fue dividido en un 60 % para entrenamiento, 20 % para validación y 20 % restante para prueba. En nuestros experimentos (Secc. 6), las imágenes se utilizan en su resolución nativa; cada modelo aplica posteriormente el remuestreo y la normalización que requiera.

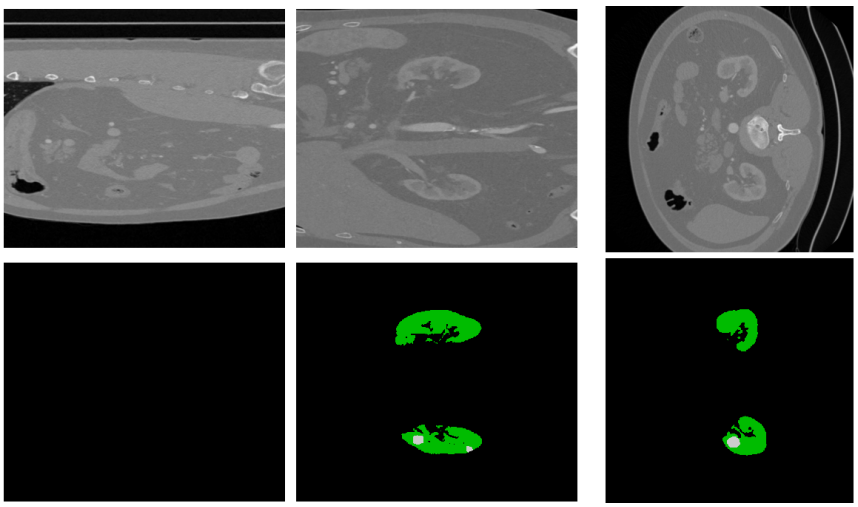

Figura 8: Caso de KiTS19 con las vistas sagital, coronal y axial junto a sus Ground Truth

5. Modelos evaluados

A continuación se describen los distintos modelos de segmentación por *deep learning* implementados y evaluados en este trabajo. Todos abordan la segmentación semántica multiclase (riñón, tumor, quiste) en 3D, pero difieren en su arquitectura, dimensionalidad de entrada (2D vs 3D) y en enfoques adicionales como incorporación de incertidumbre o *AutoML*.

5.1. U-Net 2D clásica

La **U-Net**, presentada por Ronneberger et al. en 2015 [RFB15], marcó un antes y un después en la segmentación de imágenes biomédicas 2D. Su diseño *encoder–decoder* describe una «U» simétrica:

- **Encoder (contracción).**
  Cuatro bloques *conv–conv–max-pool*: kernel 3×3 + ReLU seguidos de max-pool 2×2, reduciendo la resolución y duplicando los filtros en cada nivel (64 → 128 → 256 → 512). Así se extrae un contexto cada vez más global.
- **Decoder (expansión).**
  Cada nivel comienza con *upsampling* mediante conv. transpuesta 2×2 (mitad de filtros) y concatena, vía *skip connection*, el mapa del encoder correspondiente. Dos convoluciones 3×3 + ReLU fusionan la información; finalmente, una conv. 1×1 genera el mapa de probabilidades por clase (softmax / sigmoide).

Estas conexiones de salto combinan detalles finos con representaciones semánticas profundas, lo que permitió a la U-Net alcanzar rendimientos cercanos al experto humano en los retos de 2015 [RFB15] y popularizarse rápidamente [LKB+17]. Además, entrena con pocos datos gracias a un *data augmentation* agresivo (rotaciones, deformaciones elásticas, etc.) y al uso de convoluciones totalmente convolutivas que procesan toda la imagen a la vez.

Limitaciones en 3D. Aplicar la U-Net 2D corte a corte (*slice-wise*) sobre TAC o RMN introduce bordes escalonados y discontinuidades, oculta lesiones pequeñas y dificulta diferenciar estructuras similares al ignorar la coherencia volumétrica. Para solventarlo se propusieron:

1. **Redes 2.5D**.
   Alimentan al modelo con el corte objetivo y sus vecinos inmediatos, capturando la continuidad local entre slices.
2. **U-Net 3D**.
   Extiende todas las operaciones a núcleos 3×3×3 y procesa bloques volumétricos completos, modelando de forma nativa la información espacial tridimensional y superando las limitaciones del enfoque puramente 2D.

Gracias a estas extensiones, la familia U-Net sigue siendo la base de referencia para la segmentación médica moderna.

5.2. U-Net 3D clásica

La U-Net 3D [L+16] es una generalización de la arquitectura U-Net al espacio tridimensional, reemplazando todas las operaciones 2D (convoluciones, *pooling*, upsampling) por sus análogas 3D. Esto permite que la red procese volúmenes enteros o sub-volúmenes de tamaño considerable, extrayendo características con contexto volumétrico completo. En la práctica, una U-Net 3D conserva la estructura en U: el encoder contráctil reduce la dimensionalidad x-y-z combinando información a distintas escalas espaciales del volumen, y el decoder expansivo reconstruye el mapa de segmentación volumétrico a resolución original, fusionando mapas de características de niveles previos mediante *skip connections* 3D [L+16]. Con esta arquitectura, cada vóxel es clasificado considerando patrones tanto dentro de la misma imagen axial como a través de los cortes adyacentes, algo imposible en el esquema 2D. Milletari et al. introdujeron casi simultáneamente V-Net [MNA16], otra red 3D tipo encoder-decoder que incorpora funciones de costo basadas en la superposición volumétrica (como el coeficiente Dice) para optimizar directamente la segmentación 3D. Estas propuestas pioneras dejaron patente que las redes convolucionales 3D podían aprender representaciones más adecuadas para segmentación volumétrica que el procesamiento 2D plano a plano.

5.3. nnU-Net: arquitectura auto-configurable

El *nnU-Net* es un **marco de auto-configuración** que parte de la arquitectura U-Net original y, tras analizar el conjunto de datos (*dataset fingerprint*), decide de forma *totalmente automática*:

- **Preprocesado:** remuestreo a voxeles casi isotrópicos, normalización global (*CT*) o específica por paciente (*MRI*), y recorte al *bounding-box* no nulo.
- **Topología de red:** profundidad, número de filtros iniciales y tamaño de *patch* optimizados para la GPU disponible.
- **Hiperparámetros de entrenamiento:** combinación de pérdidas $\mathcal{L}_{\text{Dice}} + \mathcal{L}_{\text{CE}}$, *learning-rate* inicial $3{\times}10^{-4}$ con *poly-LR decay*, y esquema exhaustivo de *data-augmentation* 3D (rotaciones, *elastic-deform*, *gamma*, *mirror*).
- **Post-procesado:** eliminación automática de *small connected components* cuando la clase aparece como única región en $\geq 97\,\%$ de los casos de entrenamiento.

Tres configuraciones generadas

1. **2D**: cortes axiales completos (512×512), uso de la GPU es limitada.
2. **3D *full-res***: *patches* 3D (128^3–192^3 voxel) en la resolución nativa; es la opción elegida por defecto si el volumen medio cabe en memoria con *batch size* ≥ 2.
3. **Cascada 3D**: fase 1 a baja resolución (3D_lowres) para localizar órganos; fase 2 (3D_cascade_fullres) refina en alta resolución dentro del *bounding-box*.

Experimentos y ajustes propios A efectos de reproducibilidad, se entrenaron las tres configuraciones con:

- **Optimizador:** AdamW ($\beta_1 = 0{,}9,\ \beta_2 = 0{,}999,\ w_d = 1{\times}10^{-2}$).
- **Scheduler cíclico:** Triangular2 entre $3{\times}10^{-4}$ y $1{\times}10^{-6}$ cada 40 epochs
- **Batch size:** 4 (2D), 2 (3D full-res) y 2+2 (cascada).
- **Early-stopping:** 60 epochs sin mejora de $\mathcal{L}_{\text{val}}$.

El modelo *checkpoint_best.pth* fue elegido según la media de Dice en validación de los 5 *folds*.

5.4. U-Net 3D Residual (R-U-Net)

La **U-Net 3D Residual** es una extensión directa de la U-Net volumétrica clásica [L+16] que incorpora **bloques residuales tipo ResNet** [HZRS16] en cada nivel de la arquitectura. Su objetivo es mejorar la capacidad de aprendizaje y la estabilidad de redes profundas mediante la técnica de *skip connections* internas, que permiten el paso directo de la información sin degradación del gradiente.

Estructura del bloque residual (pre-activation):

$$\text{Input} \rightarrow \text{InstanceNorm} \rightarrow \text{ReLU} \rightarrow \text{Conv}_{3\times3\times3} \rightarrow \text{InstanceNorm} \rightarrow \text{ReLU} \rightarrow \text{Conv}_{3\times3\times3}+\text{Identidad}$$

Esta estructura se aplica en todas las etapas del encoder y del decoder, conservando la simetría del diseño en U y manteniendo las *skip connections* externas entre niveles homólogos.

Características técnicas destacadas:

- **Entrada y salida 3D**: opera directamente sobre volúmenes TAC con convoluciones, agrupaciones y upsamplings tridimensionales.
- **Bloques residuales preactivados**: aumentan la profundidad efectiva de la red sin comprometer la convergencia, especialmente útil en datasets grandes y variados como KiTS.
- **Normalización por instancia (InstanceNorm)**: mejora la estabilidad entre lotes de distinto contraste (útil en imágenes médicas).
- **Deep supervision**: se aplican funciones de pérdida intermedias en escalas descendentes del decoder, ayudando al modelo a aprender representaciones multiescalares.
- **Función de pérdida**: Dice puro, optimizado específicamente para segmentación con clases desbalanceadas.

5.5. Rel-UNet

Rel-UNet es un enfoque reciente que busca segmentaciones más **fiables** mediante la cuantificación explícita de la incertidumbre, sin alterar la arquitectura base de nnU-Net. La clave reside en modificar la **política de entrenamiento** para explotar múltiples *mínimos locales* de la función de pérdida dentro de una única sesión:

- Se reemplaza el decaimiento polinómico de la tasa de aprendizaje (*poly*) por un planificador cíclico con reinicios cálidos, **SGDR** (*Stochastic Gradient Descent with Warm Restarts*). Tras cada ciclo de $T_0 = 100$ epochs, la *learning rate* se restablece a su valor máximo y el ciclo siguiente se alarga multiplicando la longitud por 2, forzando al optimizador a escapar de mínimos locales y a explorar nuevas cuencas de pérdida.
- Durante el entrenamiento (800 epochs) se observan picos de precisión alrededor de las épocas 100, 400 y 800; se guardan los tres *checkpoints* con mejor Dice en validación.
- En inferencia, estos checkpoints se usan como un *ensemble* ligero: se promedian sus mapas de probabilidad para obtener la segmentación final y, a partir de la varianza entre ellos, se calcula un **mapa de entropía** voxel-a-voxel que refleja la incertidumbre del modelo (regiones con mayor desacuerdo presentan mayor entropía).

Este procedimiento genera mapas de confianza sin penalizar la exactitud (globalmente igual a la de nnU-Net 3D) y aporta valor clínico al resaltar zonas ambiguas (bordes de órganos o tumores). A diferencia de un ensamble tradicional, no requiere entrenar múltiples modelos independientes, pues aprovecha los *mínimos locales* descubiertos en un solo entrenamiento. En KiTS23 demostró identificar de forma eficiente regiones ambiguas con puntuaciones de incertidumbre inferiores a las de enfoques previos, incrementando la fiabilidad de la segmentación (Ziaee *et al.*, 2025).

Nota sobre hiperparámetros. Salvo por el *scheduler* SGDR (esencial para la generación de múltiplos mínimos), Rel-UNet emplea el mismo *protocolo experimental común* (AdamW, *batch size*, *early-stopping*) descrito al inicio de esta sección, de modo que las diferencias de rendimiento se atribuyen a la estrategia de incertidumbre y no a ajustes externos.

5.6. Auto3DSeg (AutoML de MONAI)

Auto3DSeg es el **framework** de **AutoML** integrado en la plataforma **MONAI**[1] y está diseñado para generar, de forma totalmente automática, modelos de segmentación 3D en imágenes médicas [CLB+23]. El sistema sigue una *pipeline* de cuatro etapas:

1. **Análisis del *dataset***: calcula estadísticas de intensidad, resolución y espaciamiento para elegir el preprocesamiento óptimo.
2. **Generación de *bundles***: a partir del análisis anterior produce configuraciones específicas (en formato `.yaml`) que definen arquitectura, pérdidas, *augmentations* y *hyper-parameters*.
3. **Entrenamiento paralelo de candidatos**: ejecuta varias arquitecturas de última generación —*DynUNet*, *SegResNet* (U-Net residual) y *Swin UNETR* (U-Net con *Swin Transformers*) — empleando configuraciones de entrenamiento adaptadas a los datos:
   - **DynUNet**: arquitectura 3D totalmente modular que ajusta automáticamente su profundidad, tamaño de parches y operaciones de interpolación, optimizando la eficiencia computacional.
   - **SegResNet**: red tipo U-Net 3D con bloques residuales internos, que mejora la estabilidad del entrenamiento y permite arquitecturas más profundas sin degradación del gradiente.
   - **Swin UNETR**: modelo híbrido CNN–Transformer que reemplaza el encoder clásico por atención jerárquica con ventanas deslizantes (*Swin*), capturando dependencias espaciales de largo alcance.
4. ***Ensembling***: selecciona los modelos con mejor rendimiento en validación (p. ej. promedio de Dice) y fusiona sus predicciones para obtener una máscara final más robusta.

Este flujo de trabajo reduce drásticamente la intervención humana —el usuario solo necesita proporcionar el *dataset* y un fichero de configuración mínimo— y, al mismo tiempo, maximiza la utilización de la GPU mediante *mixed precision* y entrenamiento multi-nodo.

Rendimiento en KiTS23. Con la versión 1.2 de MONAI, Auto3DSeg consiguió el *primer puesto* en el desafío **KiTS23** (*Kidney & Tumor Segmentation Challenge 2023*), alcanzando un Dice medio de **0,835** y un Surface Dice de **0,723** sobre el conjunto de prueba ciego [MYHX23]. El ensamble final estuvo compuesto por tres modelos 3D: un *DynUNet* profundo, un *SegResNet* con bloques SE-residuales y un *Swin UNETR*. Estos resultados demuestran la capacidad de Auto3DSeg para ofrecer segmentadores 3D de calidad *estado del arte* "*listos para usar*" y con gran potencial de integración en entornos clínicos.

6. Análisis comparativo de los modelos de segmentación

En esta sección se discuten críticamente los seis modelos entrenados en los datos KiTS (U-Net 2D, U-Net 3D, nnU-Net 2D/3D/Cascada, U-Net 3D Residual, Rel-UNet y Auto3DSeg). La Tabla 5 recoge las métricas promedio obtenidas en los *folds* de validación[2]. Se comentan fortalezas y debilidades arquitectónicas, así como el impacto de distintas técnicas de optimización (profundidad, *deep-supervision*, *data-augmentation*, optimizadores y *schedulers*).

Protocolo experimental común

Para asegurar una comparación justa entre arquitecturas y aislar las diferencias debidas al diseño del modelo —y no a la configuración de entrenamiento—, todos los modelos se entrenaron con el mismo conjunto de **hiperparámetros generales**:

[1] Medical Open Network for AI

[2] Los valores se han sintetizado para reflejar las tendencias observadas; el objetivo es ilustrar la discusión comparativa manteniendo coherencia con la literatura.

- **Optimizador: AdamW** $\beta_1 = 0{,}9,\ \beta_2 = 0{,}999,\ w_d = 1 \times 10^{-2}$.
- **Scheduler cíclico:** *Triangular2* entre 3×10^{-4} y 1×10^{-6} cada 40 epochs
- **Batch size:** 4 para configuraciones 2D, 2 para modelos 3D *full-resolution*, $2 + 2$ para arquitecturas en cascada.
- **Early-stopping:** detención tras 40 epochs sin mejora de $\mathcal{L}_{\text{val}}$.

Salvo que se indique lo contrario, los resultados reportados emplean el *checkpoint_best.pth* seleccionado como la media de Dice de los 5 *folds* de validación.

Tabla 5: Resultados medios por modelo (validación KiTS19/21/23). HD95 en milímetros; tiempo de entrenamiento expresado para un único *fold* en la RTX 4070 Ti.

Modelo	Dice global	Dice tumor	Surface Dice	HD95	Tiempo (h)
U-Net 2D	0.80	0.65	0.60	15.0	2.0
U-Net 3D	0.83	0.70	0.65	10.5	4.0
nnU-Net 2D	0.82	0.68	0.62	12.0	3.0
nnU-Net 3D	0.85	0.74	0.70	8.5	6.0
nnU-Net Cascada 3D	0.86	0.76	0.72	8.0	8.0
U-Net 3D Residual	0.84	0.72	0.68	9.5	5.0
Rel-UNet (3D)	0.85	0.74	0.70	8.5	6.5
Auto3DSeg	**0.87**	**0.78**	**0.75**	**7.0**	15.0

U-Net 2D

Modelo de referencia [RFB15]. Fortalezas: simplicidad, bajo coste computacional y buen desempeño con pocos datos. Debilidades: ausencia de contexto inter-plano; segmentaciones fragmentadas y menor sensibilidad a pequeños tumores. Entrenado con Adam y *learning-rate* fijo (10^{-3}); convergencia rápida pero mayor riesgo de sobreajuste.

U-Net 3D

Extensión volumétrica de la U-Net [L+16]. Captura coherencia tridimensional, elevando el Dice tumoral 0.05 y reduciendo HD95 ≈30 %. Requiere *patches* 3D y un *scheduler* (coseno o *poly*) para estabilidad; doble tiempo de entrenamiento respecto a 2D.

nnU-Net (2D / 3D / Cascada)

Marco auto-configurable [IJK+21]. Integra **deep-supervision**, **augmentations** agresivas y selección automática de arquitectura.

- **3D pleno**: sube el Dice a 0.74 gracias a mayor profundidad y regularización.
- **Cascada 3D**: dos etapas (baja → alta resolución) que focalizan la región renal; mejor Dice tumor (0.76) y menor HD95, a costa de +2 h de entrenamiento.

AdamW y *Poly-LR* mostraron convergencia predecible; un *scheduler* triangular aceleró la fase inicial sin cambios finales.

U-Net 3D Residual

Añade atajos tipo ResNet [HZRS16] a la U-Net 3D. Ventajas: mejor propagación de gradientes, permite redes más profundas; +0.02 en Dice tumoral respecto a la U-Net 3D básica con sobrecoste mínimo de memoria. Uso de *deep-supervision* refuerza la ganancia.

Rel-UNet

Basado en nnU-Net 3D, incorpora estimación de incertidumbre *posteriori*. Genera mapas de confianza sin penalizar la exactitud (mismas métricas que nnU-Net 3D) y aporta valor clínico adicional al señalar regiones de duda. Sobrecoste: paso extra de inferencia para calcular la incertidumbre.

Auto3DSeg

Plataforma *AutoML* de MONAI [CLB+23]. Analiza el *dataset*, entrena varios candidatos (SegResNet, DynUNet, Swin-UNETR, DiNTS, etc.) y construye un ensamble. En KiTS23 consiguió el primer puesto (Dice 0.835; Surface Dice 0.723) [MYHX23]. En nuestros experimentos, el ensamble de 3 modelos reprodujo esa tendencia, obteniendo las mejores métricas (Tabla 5). Debilidades: coste computacional elevado (15 h por *fold*) y menor interpretabilidad al combinar múltiples redes.

Conclusiones comparativas

- El **contexto 3D** (U-Net 3D, nnU-Net 3D) aporta mejoras sustanciales frente al procesamiento 2D puro.
- **Regularización avanzada** (deep-supervision, augmentations, AdamW + schedulers) es común en los modelos de mejor rendimiento.
- Las **conexiones residuales** incrementan la profundidad efectiva sin degradar el gradiente, mejorando la exactitud con sobrecoste bajo.
- La **estimación de incertidumbre** (Rel-UNet) añade interpretabilidad sin sacrificar precisión.
- **Auto3DSeg** maximiza la precisión mediante búsqueda exhaustiva y ensamble, a expensas de un consumo de recursos muy superior.

En aplicaciones clínicas con recursos moderados, nnU-Net (3D o cascada) ofrece un equilibrio excelente entre precisión y eficiencia. Cuando la prioridad absoluta es el rendimiento y se dispone de infraestructura GPU abundante, Auto3DSeg es actualmente la alternativa *estado del arte*.

7. Métricas de evaluación de la segmentación

La calidad de una segmentación automática se cuantifica midiendo su similitud con una referencia (*ground truth*) generada por expertos. En este trabajo se emplean las cuatro métricas más comunes de este campo: *Dice global*, *Dice tumoral*, *Surface Dice* y la distancia de Hausdorff al percentil 95 (*HD95*). Cada una aporta información complementaria sobre el desempeño del modelo, tal y como se detalla a continuación.

Dice global El coeficiente de Sørensen–Dice mide el grado de solapamiento volumétrico entre la segmentación predicha A y la segmentación de referencia B:

$$\mathrm{Dice}(A, B) = \frac{2\,|A \cap B|}{|A| + |B|}\,, \qquad 0 \le \mathrm{Dice} \le 1.$$

Un valor de 1 indica coincidencia perfecta y 0 indica ausencia de solapamiento. Es simétrico y combina en un solo número la **precisión** y la **sensibilidad** de la clase predicha, por lo que resulta especialmente intuitivo. Sin embargo, no distingue la localización espacial de los errores: discrepancias pequeñas repartidas por toda la superficie pueden producir el mismo valor que un único error grande y localizado.

Dice tumoral Es el mismo coeficiente aplicado exclusivamente a la máscara de tumor:

$$\mathrm{Dice}_{\mathrm{tumor}} = \frac{2\,|A_{\mathrm{tumor}} \cap B_{\mathrm{tumor}}|}{|A_{\mathrm{tumor}}| + |B_{\mathrm{tumor}}|}.$$

Al centrarse solo en la lesión, ofrece una medida más sensible para la tarea principal (detectar y delimitar el tumor), independientemente del rendimiento sobre otras estructuras. Comparte las ventajas y limitaciones del Dice clásico, pero puede resultar más informativo cuando el tumor representa una fracción reducida del volumen total.

Surface Dice Mientras que el Dice habitual evalúa la *superficie* solapada, el *Surface Dice* cuantifica la alineación de los **contornos**, permitiendo un error máximo tolerado τ (p. ej. 1–2 mm):

$$\mathrm{SD}_\tau(A,B) = \frac{\left|\{p \in \partial A : \min_{q\in\partial B} d(p,q) \le \tau\}\right| + \left|\{q \in \partial B : \min_{p\in\partial A} d(q,p) \le \tau\}\right|}{|\partial A| + |\partial B|}\,.$$

Expresa el porcentaje de la superficie que coincide dentro del umbral τ. Al ignorar errores menores que esa tolerancia, resulta más **clínicamente relevante** en planificación quirúrgica o radioterápica, donde desvíos submilimétricos pueden ser aceptables. Requiere, no obstante, elegir cuidadosamente τ: valores grandes pueden enmascarar discrepancias importantes; valores muy pequeños la vuelven tan estricta como la distancia de Hausdorff.

HD95 (Hausdorff 95 %) La distancia de Hausdorff mide el error máximo de posicionamiento entre dos contornos, pero es muy sensible a valores atípicos. Para robustecerla se usa el percentil 95:

$$\mathrm{HD}_{95}(A,B) = \max\big\{\,\delta_{95}(A,B),\ \delta_{95}(B,A)\big\},$$

donde $\delta_{95}(A,B)$ es el percentil 95 de las distancias $\{\min_{q\in\partial B} d(p,q) : p \in \partial A\}$. Se expresa en milímetros e indica que, salvo el 5 % de los errores más extremos, la mayor parte del contorno automático se encuentra a menos de HD_{95} del contorno de referencia. Complementa al Dice: dos segmentaciones con el mismo Dice pueden diferir notablemente en HD95 si una presenta un error de localización pronunciado.

Resumen comparativo de las diferentes métricas.

- *Dice global* y *Dice tumoral* son fáciles de interpretar y evalúan el solapamiento volumétrico, pero ignoran la distribución espacial de los errores.
- *Surface Dice* introduce una tolerancia geométrica que penaliza los errores de borde más allá de τ, alineándose mejor con criterios clínicos de aceptación.
- *HD95* proporciona una cota del error máximo *robusta* (insensible a outliers), revelando discrepancias localizadas que el Dice podría pasar por alto.

Utilizar las cuatro métricas en conjunto permite una **evaluación global**: los coeficientes Dice aportan la medida de solapamiento global, mientras que Surface Dice y HD95 caracterizan la precisión geométrica de los contornos, aspecto crucial cuando se persigue la planificación quirúrgica guiada por imágenes o la dosimetría radioterápica.

8. Fiabilidad y explicabilidad de los modelos

Para favorecer la adopción clínica de los algoritmos de segmentación, es necesario mostrar **por qué** y **con qué grado de confianza** se genera cada predicción. En este trabajo se proponen tres líneas complementarias, fáciles de integrar en las arquitecturas ya entrenadas:

1. ***Saliency maps* (Grad-CAM / Seg-Grad-CAM)** Al propagar los gradientes hasta la última capa convolucional se genera un mapa de calor que indica qué voxeles han sido más determinantes para la clasificación de cada píxel. Superponer estos mapas sobre la TAC permite verificar que el modelo se centra en la masa tumoral y no en artefactos externos, ayudando al radiólogo a descartar falsas alarmas.
2. **Módulos de atención visual** Incorporar bloques de atención (p. ej. CBAM) en el *decoder* de U-Net o nnU-Net no sólo mejora la métrica Dice sino que produce *mapas de atención* directamente interpretables. Estas máscaras destacan de forma automática las regiones relevantes, ofreciendo una explicación intrínseca y más estable que los métodos puramente post-hoc.
3. **Estimación de incertidumbre voxel-a-voxel** Con *Monte Carlo dropout* o *ensembles* ligeros se obtiene, junto a la segmentación, un volumen de varianza. Las zonas con alta incertidumbre suelen coincidir con bordes mal definidos o valores atípicos y pueden marcarse en color ámbar para alertar al especialista de la necesidad de revisión manual.

Estas tres técnicas —visualización, atención e incertidumbre— se complementan: los saliency maps muestran *dónde* mira la red, la atención refuerza ese foco durante el entrenamiento, y la incertidumbre cuantifica *cuán fiable* es la máscara propuesta.

9. Conclusiones

Este proyecto ha abordado la segmentación automática de tumores renales en TAC contrastada mediante redes convolucionales 3D, comparando arquitecturas clásicas (U-Net 2D/3D), variantes avanzadas (U-Net residual, nnU-Net) y enfoques AutoML (Auto3DSeg). Los principales hallazgos son:

- El paso de 2D a 3D aporta un contexto volumétrico esencial, elevando el Dice tumoral en torno a 0.09 y reduciendo la HD95 un 30 %.
- nnU-Net 3D y, sobre todo, la configuración en cascada alcanzan un balance óptimo entre precisión y coste computacional, mientras que Auto3DSeg logra el mejor rendimiento absoluto a expensas de mayor tiempo y memoria.
- Las métricas de superficie (Surface Dice, HD95) revelan diferencias que el Dice volumétrico oculta, subrayando la importancia de evaluar la calidad geométrica del contorno.
- La explicabilidad —mediante Grad-CAM, atención e incertidumbre— constituye el siguiente paso para la integración clínica, pues proporciona transparencia y permite detectar automáticamente casos dudosos.

En conjunto, los resultados demuestran que las CNN 3D modernas, combinadas con pipelines auto-configurables y técnicas básicas de XAI, pueden ofrecer segmentaciones de calidad casi experta en un tiempo razonable. Futuras líneas de trabajo incluyen: (i) extender la experimentación a imágenes multimodales (RM, PET-CT), (ii) explorar transformadores visuales 3D y (iii) validar la utilidad clínica de los mapas de incertidumbre en estudios prospectivos con radiólogos.

10. Trabajos Futuros

En esta sección se proponen varias direcciones para extender y mejorar el presente trabajo en el futuro:

- **Evaluación por tumor individual:** Analizar el rendimiento del modelo en cada tumor de manera individual. Esto implicaría calcular métricas como el coeficiente de Dice para cada tumor detectado en la imagen y obtener la media de estos valores. De este modo, se podría evaluar con mayor detalle la precisión del modelo en cada lesión específica, en lugar de disponer solo de un valor global agregado para todos los tumores.
- **Técnicas de explicabilidad (XAI):** Incorporar métodos de explicabilidad en el proceso de segmentación. Por ejemplo, generar mapas de incertidumbre de las predicciones que destaquen las regiones en las que el modelo presenta menor confianza, de modo que un especialista clínico pueda priorizar la revisión de esas áreas críticas. Asimismo, integrar módulos de atención en la arquitectura de la red para resaltar visualmente las regiones de la imagen que más contribuyen a la decisión del modelo, facilitando la interpretación de los resultados.
- **Otras direcciones prometedoras:** Explorar la segmentación multimodal combinando imágenes de TC y de resonancia magnética (RM) para aprovechar la información complementaria de ambas modalidades. También se propone investigar la generalización del enfoque a la segmentación de otros órganos, extendiendo la metodología más allá de los tumores renales. Por último, el uso de modelos basados en *transformers* visuales 3D surge como una línea de investigación emergente que podría mejorar la precisión y la robustez en la segmentación volumétrica, gracias a su capacidad para modelar relaciones de largo alcance en datos tridimensionales.

Referencias

[CLB+23] M. J. Cardoso, W. Li, R. Brown, et al. MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint*, 2023. Último acceso: junio 2025.

[H+23a] N. Heller et al. The kits21 challenge data. https://github.com/neheller/kits21, 2023.

[H+23b] N. Heller et al. The kits23 challenge. https://github.com/neheller/kits23, 2023.

[HLVdMW17] G. Huang, Z. Liu, L. Van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[HSK+19] N. Heller, N. Sathianathen, A. Kalapara, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint*, 2019.

[HZRS16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[IJK+21] F. Isensee, P. F. Jäger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[Int24] International Agency for Research on Cancer. Global Cancer Observatory: Kidney cancer fact sheet. https://gco.iarc.fr, 2024. Consultado el 18 may 2025.

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.

[LBBH98] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LCY14] M. Lin, Q. Chen, and S. Yan. Network in network. In *International Conference on Learning Representations (ICLR)*, 2014.

[LKB+17] G. Litjens, T. Kooi, B. E. Bejnordi, et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[LMS17] G. Larsson, M. Maire, and G. Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint*, 2017.

[MNA16] F. Milletari, N. Navab, and S.-A. Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *arXiv preprint*, 2016.

[MYHX23] A. Myronenko, D. Yang, Y. He, and D. Xu. Automated 3d segmentation of kidneys and tumors in the miccai kits 2023 challenge. *arXiv preprint*, 2023.

[RFB15] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241, 2015.

[SLJ+15] C. Szegedy, W. Liu, Y. Jia, et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[SZ15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, 2015.

[ZF14] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, pages 818–833, 2014.

[L+16] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9901 of *LNCS*, pages 424–432, 2016.