UNIVERSIDAD DE VALLADOLID

FACULTAD DE MEDICINA

ESCUELA DE INGENIERÍAS INDUSTRIALES

TRABAJO DE FIN DE GRADO

GRADO EN INGENIERÍA BIOMÉDICA

Detección automática de estructuras de la vía aérea en imágenes de broncoscopia mediante técnicas de aprendizaje profundo

Autor:

D. Alejandro Medina Diez

Tutor:

D. Daniel Álvarez González

D. Tomás Ruiz Albi

Valladolid, 30 de julio de 2025

TÍTULO:	**Detección automática de estructuras de la vía aérea en imágenes de broncoscopia mediante técnicas de aprendizaje profundo**
AUTOR:	**D. Alejandro Medina Diez**
TUTORES:	**D. Daniel Álvarez González** **D. Tomás Ruiz Albi**
DEPARTAMENTO:	**Departamento de** **Teoría de la Señal y Comunicaciones e Ingeniería Telemática** **Medicina, Dermatología y Toxicología**

TRIBUNAL

PRESIDENTE:	**D. Mario Martínez Zarzuela**
SECRETARIO:	**D. Daniel Álvarez González**
VOCAL:	**D. Tomás Ruiz Albi**
SUPLENTE 1:	**D.ª María García Gadañón**
SUPLENTE 2:	**D. Gonzalo C. Gutiérrez Tobal**

FECHA: 30/07/2025

CALIFICACIÓN:

A todos los pacientes.

AGRADECIMIENTOS

En primer lugar, me gustaría agradecer a todos los pacientes que se han prestado a ceder sus imágenes de broncoscopia para que yo pudiera realizar este trabajo. Porque han sido generosos, muchas veces incluso en momentos difíciles para ellos y sus familias.

Gracias al Dr. Milko Terranova, por la ayuda recibida y por acceder tan gustosamente a ayudarme en todo lo que fuera necesario para este trabajo y a recopilar los vídeos de broncoscopia solucionando cualquier problema que pudiera surgir. Gracias también al Dr. Tomás Ruiz Albi por facilitarme la entrada al Hospital Universitario Río Hortega.

Gracias a mi familia. A mi madre y a mi padre por el apoyo, los consejos y hacerme sentir querido. A mi hermano Javier por inspirarme a esforzarme, ser mejor persona y demostrarme que el tiempo es solo una limitación cuando se pretende dormir.

Gracias a mis amigos, por acompañarme durante estos años. Por hacerme sentir parte de algo grande y por darme la seguridad de que la vida a la que aspiro es posible. Gracias a los amigos que me ha regalado la carrera y gracias a aquellos que lo llevan siendo desde que tengo uso de razón. Gracias en especial a Celia, por su apoyo incondicional.

Por último, me gustaría agradecer profundamente a mi tutor, el Dr. Daniel Álvarez González por muchos motivos. Primero por ser un gran profesor, capaz de transmitir e inspirar por medio de la enseñanza y la cercanía. Por confiar en mí, en mis ideas y empujarme a lograr la excelencia académica. Gracias por tus buenas palabras, por guiarme y ofrecerme tu ayuda. Finalmente, gracias por todo el apoyo recibido durante el desarrollo de este trabajo, desde septiembre.

Resumen

Antecedentes. La broncoscopia es una técnica endoscópica muy utilizada durante la práctica clínica que permite visualizar las vías del sistema respiratorio y acceder físicamente al árbol bronquial con múltiples fines clínicos o diagnósticos. La responsabilidad de localizar la sonda del broncoscopio durante la intervención recae únicamente en el neumólogo que realiza la prueba, dependiendo exclusivamente de su conocimiento y experiencia, si bien existen algunos métodos de imagen tradicionales basados en radiación ionizante que ayudan en esta tarea. Actualmente se encuentran en desarrollo diversos métodos para localizar la sonda del broncoscopio durante la intervención sin efectos secundarios y de altas prestaciones, entre los que destacan la broncoscopia asistida robóticamente, la broncoscopia por navegación electromagnética y la asistencia mediante algoritmos de inteligencia artificial. Sin embargo, estos métodos todavía no se han establecido en la práctica clínica.

Hipótesis y Objetivos. El presente trabajo se ha desarrollado bajo la hipótesis de que es posible localizar la sonda de broncoscopia a partir únicamente de imágenes de broncoscopia segmentadas, inexplorado hasta el momento por la comunidad científica. El objetivo general del estudio consistió en diseñar y validar un modelo predictivo capaz de detectar y segmentar automáticamente las estructuras naturales de la vía aérea hasta la primera bifurcación (bronquio izquierdo, LB; bronquio derecho, RB; carina traqueal, CT). Para conseguirlo se han establecido varios objetivos específicos, entre los que destacan (i) la evaluación de distintas arquitecturas y algoritmos de DL para determinar cuál alcanza el mayor rendimiento predictivo en esta tarea y (ii) la creación de un sistema de localización de la sonda del broncoscopio basado en imágenes segmentadas de broncoscopia.

Materiales y Métodos. Para este trabajo se recopilaron un total de 35 vídeos de manera prospectiva en el Hospital Universitario Río Hortega, de pacientes sin alteraciones anatómicas relevantes de las vías respiratorias. Estos vídeos se han dividido en una proporción de 55%-45% entre conjunto de entrenamiento (19) y test (16), con una contribución media de 245 imágenes por paciente. Posteriormente, los vídeos fueron preprocesados y etiquetados manualmente creando un *dataset* de las imágenes y sus

respectivas etiquetas, que cuenta con 8591 imágenes; 4787 imágenes para el conjunto de entrenamiento y 4164 para el conjunto de test. Gracias al *framework* nnU-Net se han creado y comparado dos arquitecturas distintas, Plain U-Net y ResEnc U-Net, basadas en la U-Net original. Finalmente, se ha creado un sistema de localización a partir del cual, por medio de la identificación de las etiquetas presentes en una imagen segmentada, se le asocia una localización específica dentro del árbol bronquial.

Resultados. Para evaluar los resultados de las arquitecturas de segmentación se han empleado varias métricas. Las más relevantes son la *Intersection over Union* (IoU) y el *Dice Similarity Coefficient* (DSC) promedios. La arquitectura Plain U-Net alcanzó valores de 0.69 para el IoU (0.72 LB, 0.73 RB, 0.62 CT) y de 0.814 para el DSC (0.84 LB, 0.84 RB, 0.77 CT), mientras que la arquitectura ResEnc U-Net alcanzó un IoU de 0.674 (0.71 LB, 0.73 RB, 0.59 CT) y un DSC de 0.802 (0.83 LB, 0.84 RB, 0.74 CT). Estos resultados dotan a la primera arquitectura de un mayor poder de segmentación. En relación con la tarea de localización, la arquitectura Plain U-Net alcanzó una exactitud media del 83.67%, mientras que la arquitectura ResEnc U-Net presentó una exactitud media del 91%, siendo esta última más apta para la tarea de localización. Mientras que la arquitectura Plain U-net mostró un mayor desempeño en la tarea de segmentación, la ResEnc U-Net resultó de mayor eficacia para la tarea de localización de la sonda, dotando a la herramienta de un mayor compromiso entre la precisión morfológica y la aplicabilidad clínica.

Conclusiones. Se ha creado el primer dataset de imágenes de broncoscopia segmentadas. A partir de este dataset se han creado dos herramientas de segmentación automática de imágenes de broncoscopia. Los resultados obtenidos sugieren que la localización de la sonda de broncoscopia mediante imágenes segmentadas representa una alternativa viable frente a otros métodos ampliamente consolidados, aportando ventajas en términos de rapidez, complejidad y coste.

Palabras clave

Broncoscopia, Neumología Intervencionista, Navegación Bronquial, Visión Artificial, Inteligencia Artificial, Deep Learning, Segmentación de imagen médica, U-Net, NnU-Net.

Abstract

Background. Bronchoscopy is an endoscopic technique widely used during routine clinical practice that allows visualization of the respiratory tract and physical access to the bronchial tree for multiple clinical or diagnostic purposes. The responsibility for locating the bronchoscope probe during the procedure lies solely with the pulmonologist performing the examination, relying exclusively on their knowledge and experience, although there are some traditional imaging methods based on ionizing radiation that assist clinicians in this task. Currently, several methods are under development to locate the bronchoscope probe during the procedure without side effects and with high performance, among which robotic assisted bronchoscopy, electromagnetic navigation bronchoscopy, and assistance through artificial intelligence algorithms. However, these methods have not yet been established in clinical practice.

Hypothesis and Objectives. The present work was developed under the hypothesis that it is possible to locate the bronchoscopy probe using only segmented bronchoscopy images, an approach unexplored so far by the scientific community. The main objective of the study was to design and validate a predictive model capable of automatically detecting and segmenting the natural structures of the airway up to the first bifurcation (left bronchus, LB; right bronchus, RB; and tracheal carina, CT). To achieve this goal, several specific objectives were also established, including (i) the evaluation of different architectures and deep learning algorithms, to determine which is better suited for this task, and (ii) the creation of a bronchoscope probe localization system based on segmented bronchoscopy images.

Materials and Methods. For this study, a total of 35 videos were prospectively collected at the Hospital Universitario Río Hortega, from patients without relevant anatomical alterations of the respiratory tract. These videos were divided in a 55%-45% proportion into training set (19) and test set (16). They were then manually preprocessed and labelled, creating a dataset of images and their corresponding labels. Thanks to the nnU-Net framework, two different architectures were created and compared: Plain U-Net and ResEnc U-Net, both based on the original U-Net architecture. Finally, a localization

system was created through which, by identifying the labels present in a segmented image, a specific location within the bronchial tree is assigned.

Results. To evaluate the results of the segmentation architectures, several metrics were used. The most relevant are the average Intersection over Union (IoU) and Dice Similarity Coefficient (DSC). The Plain U-Net architecture showed values of 0.69 for IoU (0.72 LB, 0.73 RB, 0.62 CT) and 0.814 for DSC (0.84 LB, 0.84 RB, 0.77 CT), while the ResEnc U-Net architecture showed an IoU of 0.674 (0.71 LB, 0.73 RB, 0.59 CT) and a DSC of 0.802 (0.83 LB, 0.84 RB, 0.74 CT). These results give the former architecture greater segmentation power. Regarding the localization task, the Plain U-Net architecture achieved an average accuracy of 83.67%, while the ResEnc U-Net architecture showed an average accuracy of 91%, making the latter more suitable for the localization task. While the Plain U-Net architecture performs better in the segmentation task, the ResEnc U-Net proved to be more effective for locating the probe, providing the tool with a better balance between morphological precision and clinical applicability.

Conclusions. The first dataset of segmented bronchoscopy images ever has been created. Based on this dataset, two automatic segmentation tools for bronchoscopy images were developed. The results suggest that bronchoscope probe localization through segmented images represents a feasible alternative to other widely established methods, offering advantages in speed, complexity, and cost.

Keywords

Bronchoscopy, Interventional Pulmonology, Bronchial Navigation, Computer Vision, Artificial Vision, Artificial Intelligence, Deep Learning, Medical Image Segmentation, U-Net, nnU-Net.

Índice general

Índice de figuras

Índice de tablas

Glosario de acrónimos

Adan: Adaptative Nesterov Momentum Algorithm

BCE: Binary Cross Entropy

CBCT: Cone-Beam computed tomography

CNN: Convolutional Neural Network (Red Neuronal Convolucional)

DC: Dice Loss

DL: Deep Learning

DSC: Dice Similarity Coefficient

EBUS: Endobronchial ultrasound-guided

EMA: Exponential Moving Average

ENB: Electromagnetic Navigation Bronchoscopy

EPOC: Enfermedad Pulmonar Obstructiva Crónica

FN: False Negative

FP: False Positive

FPN: Panoptic Feature Pyramid Network

GT: Ground Truth

GAN: Generative Adversarial Networks

HD95: Haussdorff Distance (95%)

HURH: Hospital Universitario Río Hortega

IA: Inteligencia Artificial

IoU: Intersection over Union

LR: Learning Rate

LSTM: Long Short-Term Memory

Mask R-CNN: Mask Region-based Convolutional Neural Networks

ML: Machine Learning

NeRF: Neural Radiance Fields

NI: Neumología Intervencionista

Nn-Unet: no-new U-Net

PPV: Positive Predictive Value

RAB: Robotic Assisted Bronchoscopy

RM: Resonancia Magnética

RNA: Red Neuronal Artificial

RNN: Redes Neuronales Recurrentes

RVA: Revisión de las vías aéreas

RX: Rayos X

TC: Tomografía Computarizada

TN: True Negative

TP: True Positive

VGG: Visual Geometry Group

1. Introducción

A lo largo de este capítulo se recoge todo el marco teórico y contexto del trabajo. Se describe brevemente lo que es la broncoscopia y su uso mayoritario en la neumología intervencionista. Posteriormente, se define el concepto de localización bronquial y se explora el estado del arte referente a localizar con exactitud el emplazamiento del broncoscopio dentro del árbol bronquial durante la propia intervención. Finalmente, se ahonda en la anatomía del árbol bronquial y en el concepto de segmentación de imagen médica.

1.1 Introducción a la broncoscopia y neumología intervencionista

La broncoscopia es un procedimiento endoscópico relativo al sistema respiratorio que permite visualizar y acceder a la vía aérea y, más específicamente, al árbol bronquial. Nace como una técnica variante de la esofagoscopia, habiendo evolucionado esta última de la laringoscopia, en el año 1876, de la mano de Gustav Killian. Acuñado como "*directe bronkoscopie*", fue utilizado para la extracción de un cuerpo extraño de las vías aéreas de un granjero gracias al uso conjunto de iluminación y succión [1].

Si bien por sí sola la broncoscopia únicamente resulta de utilidad como herramienta de exploración y diagnóstico visual, gracias a la incorporación de instrumental externo como herramientas de aspiración, biopsia, congelación, ablación e incluso vaporización, este procedimiento adopta una nueva dimensión y es capaz de abarcar desde el diagnóstico hasta el tratamiento de numerosas patologías pulmonares. Algunas de las aplicaciones más comunes incluyen la aspiración de secreciones con fines terapéuticos, la identificación y diagnóstico de lesiones en la vía aérea o la biopsia transbronquial para el diagnóstico de enfermedades pulmonares parenquimales [2]. Entre las patologías pulmonares más abordadas por esta técnica se incluyen cáncer, enfisema, asma, enfermedad pulmonar obstructiva crónica (EPOC) o anomalías de la vía aérea, como estenosis o colapso dinámico excesivo durante la espiración. De la misma forma, su utilización combinada con técnicas de imagen como la fluoroscopia, ultrasonidos o la

tomografía computarizada (TC) han mejorado la navegación y acceso a lesiones periféricas. Sin el empleo de estas técnicas de imagen, el rendimiento diagnóstico de la broncoscopia convencional es muy limitado [3].

Esta técnica endoscópica encuentra su mayor protagonismo en la neumología intervencionista (NI). Esta área, reconocida como una subespecialidad de la neumología desde el año 2001 [4], es una rama relativamente nueva de la medicina que permite un enfoque mínimamente invasivo al diagnóstico y tratamiento de patologías de la anatomía de la vía aérea, parenquimales o pleuríticas. La broncoscopia es la técnica fundamental sobre la que giran la mayoría de las aplicaciones en la NI. Las principales son: *(i)* broncoscopia diagnóstica, *(ii)* broncoscopia diagnóstica en enfermedades mediastínicas, *(iii)* broncoscopia diagnóstica en lesiones periféricas, *(iv)* resección y ablación de nódulos periféricos, *(v)* broncoscopia para el diagnóstico y manejo de obstrucciones centrales de las vías aéreas y *(vi)* válvulas endobronquiales en el tratamiento del enfisema [4], [5], [6]. Sin la broncoscopia, todas estas técnicas serían imposibles de manera mínimamente invasiva y, en caso de ser posibles de otra manera, requerirían de cirugías torácicas invasivas, más peligrosas y de menor efectividad.

El principal inconveniente de la broncoscopia es la dependencia total en los conocimientos anatómico-médicos del profesional que la utiliza para la localización dentro del árbol bronquial. Si bien esta tarea normalmente no resulta ser muy complicada, existen determinados casos en los que una localización fina es necesaria para realizar ciertos procedimientos, como pueden ser la toma de biopsias o el acceso a nódulos periféricos pequeños [1]. En los últimos años se han realizado relativamente pocas innovaciones orientadas a facilitar la labor del intervencionista que hace uso de la broncoscopia. La mayor parte de estas innovaciones están orientadas a mejorar la calidad de imagen de la cámara endoscópica y a disminuir el tamaño del equipamiento utilizado. Sin embargo, existen algunas innovaciones centradas en el manejo del instrumento o en conocer la localización exacta en árbol bronquial y facilitar la navegación por este. Dentro de los avances más recientes destacan la introducción de la broncoscopia robótica, la navegación por campos de radiancia neural y la navegación electromagnética para facilitar tanto la navegación como el tratamiento pertinente [7]. Aunque por ahora son avances recientes, están logrando mejores resultados en el diagnóstico que la broncoscopia convencional [7], [8], [9]. En la Figura 1 se pueden visualizar la realización de una ecobroncoscopia, así como la instrumentación utilizada.

Figura 1. Intervención broncoscópica realizada en el Hospital Quirónsalud Málaga por el servicio de neumología tras la incorporación de la ecobroncoscopia (EBUS).

1.2 Anatomía del árbol bronquial

El árbol bronquial es la estructura anatómica ramificada que comprende las vías aéreas desde la laringe hasta los pulmones. Está compuesta por la tráquea, los bronquios principales, los bronquios lobares, los bronquios segmentarios, los bronquiolos y los alvéolos. A nivel de broncoscopia, debido al calibre del propio broncoscopio, no se llega a acceder a los bronquiolos y los alvéolos.

El árbol bronquial posee una nomenclatura propia para discernir entre las distintas zonas y ramas anatómicas. En función de las subdivisiones, podemos clasificar el árbol bronquial según la nomenclatura internacional de Collins *et al.* y Vaz Rodrigues *et al.* [10], [11] y el capítulo de anatomía broncoscópica Atlas *Video-Atlas of VATS Pulmonary Sublobar Resections* [12] de la siguiente manera, descrita en la Tabla 1.

Tabla 1. Anatomía del árbol bronquial. Tabla adaptada de Vaz Rodrigues et al. [11]

Right Lung		***Left Lung***	
Right Upper Lobe		**Left Upper Lobe**	
B1	*apical segment*	B1 + 2	*apicoposterior segment*
B2	*posterior segment*	B3	*anterior segment*
B3	*anterior segment*	**Língula**	
Middle Lobe		B4	*superior segment*
B4	*lateral segment*	B5	*inferior segment*
B5	*medial segment*	**Left Lower Lobe**	
Right Lower Lobe		B6	*apical basal segment*
B6	*apical basal segment*	B7 + 8	*anterior basal segment*
B7	*medial basal segment*	B9	*lateral basal segment*
B8	*anterior basal segment*	B10	*posterior basal segment*
B9	*lateral basal segment*		
B10	*posterior basal segment*		

Dentro de esta clasificación, destacamos la primera división en bronquios derecho e izquierdo, asociándose cada uno con su respectivo pulmón. El bronquio derecho presenta 3 ramificaciones o bronquios lobares, asociados cada uno a su respectivo lóbulo pulmonar (superior, medio e inferior). A su vez, el bronquio izquierdo presenta 2 bronquios lobares asociados a sus respectivos lóbulos pulmonares (superior e inferior). Cada bronquio lobar se divide en sus respectivos bronquios segmentarios, cada uno asociado a su segmento pulmonar tanto por el bronquio como por la enervación y vascularización propias.

Existen numerosas variantes anatómicas, que con mayor o menor frecuencia pueden alterar la anatomía considerada como normal de los pacientes. En la revisión sistemática de las variantes anatómicas de Vaz Rodrigues *et al*. se describen un total de 20 variaciones anatómicas, con una prevalencia del 43% (79 pacientes de 207 examinados). Las más comunes son las del pulmón derecho, con un 64.7% de ocurrencia dentro de todas las variantes observadas, dejando a las del pulmón izquierdo con un 31.9%. La variante más recurrente es la bifurcación B1 + B2 con B3 [11].

En la Figura 2 se muestran algunas imágenes donde se puede apreciar la complejidad de la anatomía y todo el entramado estructural que da lugar al sistema respiratorio.

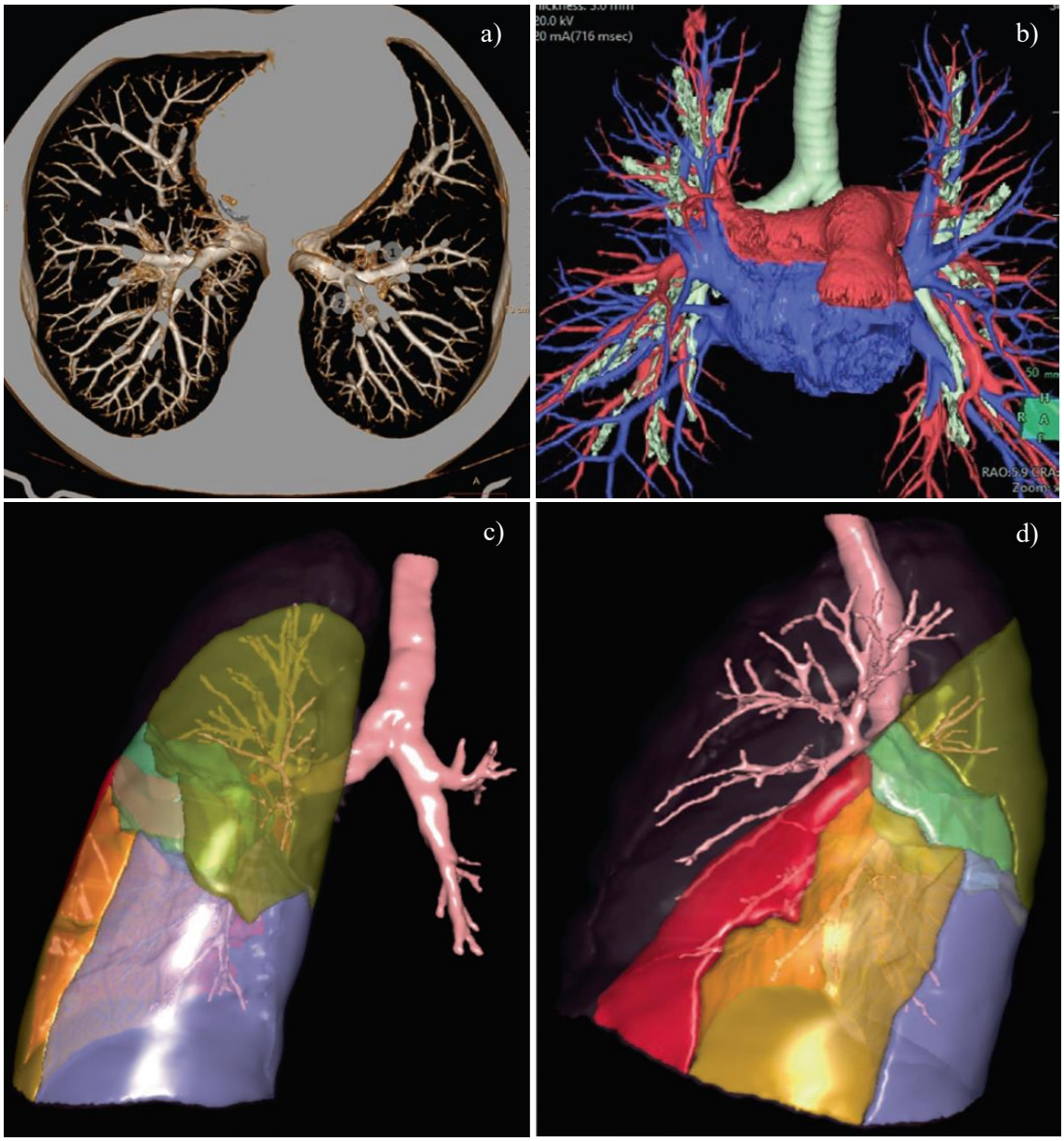

Figura 2. Ejemplos de la anatomía del árbol bronquial a partir de imágenes de TC. Las estructuras son: a) superposición de vías aéreas en TC; b) vasos sanguíneos asociados al sistema respiratorio; c) y d) pulmón izquierdo desde diferentes ángulos. Imágenes extraídas de [12].

1.3 Estado del arte en imagenología de navegación broncoscópica

La realidad clínica habitual en el área de la navegación bronquial depende de cada hospital, aunque generalmente se realiza una TC previa a la exploración. Esta prueba se puede utilizar para realizar una planificación de la broncoscopia, que generalmente es necesaria para la mayoría de los métodos de navegación bronquial existentes. Durante la propia intervención, la práctica clínica en el área de la navegación bronquial se basa en la

realización de radiografías a pie de cama durante el propio procedimiento, exponiendo a los presentes a posibles radiaciones ionizantes y la necesidad de hacer uso de mandiles plomados para su protección. Sin embargo, en los últimos años han comenzado a desarrollarse nuevas técnicas de navegación bronquial que destacan por tener una alta precisión, aunque con elevado coste y con el requerimiento añadido de necesitar una mayor infraestructura en el quirófano.

Los últimos avances en esta área de trabajo giran en torno a tres ejes principales:

(i) **Imagenología híbrida y *Robotic Assisted Bronchoscopy* (RAB).** Consiste en la utilización de distintos métodos de obtención de imágenes para combinarlos y obtener una localización más precisa que la que se obtendría si únicamente se utilizara un método. Dentro de este apartado encontramos la combinación de la broncoscopia con técnicas como *Cone-beam Computed Tomography* (CBCT), fluoroscopia en tiempo real o broncoscopia guiada por ultrasonidos (EBUS). Su mayor uso y necesidad reside en la toma de biopsias transbronquiales, donde el estándar es hacer uso de ultrasonidos para localizar lesiones fuera del árbol bronquial. Esto mejora notablemente la localización de las lesiones y, por tanto, el diagnóstico [13], [14]. La fluoroscopia en tiempo real y el uso de CBCT siguen todavía en desarrollos tempranos, pero prometen resultar de mucha utilidad en el futuro [4]. La integración de todas estas técnicas está siendo posible gracias a la introducción de la RAB, que gracias a un enfoque similar a las cirugías asistidas robóticamente, es capaz de dotar a los sanitarios de un mayor control sobre la intervención, con plataformas como Monarch™, Ion™ o Galaxy™. Todas estas herramientas son capaces de incorporar sistemas de navegación por imagen como CBCT, EBUS o fluoroscopia en tiempo real [4], [15], [16].

(ii) **Asistencia a la navegación por algoritmos de inteligencia artificial.** Es la forma más novedosa y por ahora menos utilizada en la práctica clínica de la navegación bronquial. Se trata de técnicas que hacen uso de la inteligencia artificial (IA) para localizar tridimensionalmente el broncoscopio durante la propia intervención por medio de las imágenes de la prueba. Además, no es necesaria infraestructura adicional. Destacan varios algoritmos que favorecen a la localización. Se centran en el análisis de imagen para calcular parámetros como la profundidad o clasificar imágenes para detectar distintas bifurcaciones según sus lúmenes. Los avances más notorios para detectar la orientación del

broncoscopio en referencia a sus grados de libertad son DD-VNB (*Depth-based Dual-Loop Visual Navigation*) y PANS (*Probabilistic Airway Navigation System*) [17], [18]. Otros, como *Broncho Track* o la red de clasificación de Li *et al.* son capaces de detectar y clasificar lúmenes y asociarlos a posiciones de las vías aéreas logrando muy buena precisión en la localización bronquial, gracias a las redes neuronales convoluciones (CNN) [19], [20]. Finalmente, existen otro tipo de algoritmos denominados NeRF (*Neural Radiance Fields*), que mediante las imágenes del broncoscopio son capaces de localizar su posición tridimensionalmente y no solo la bifurcación en la que se encuentra [8].

(iii) **Mecanismos externos a la imagen.** Destaca el uso de imagen de TC y campos electromagnéticos. En base a estas tecnologías se han creado técnicas como BronchoX, perteneciente al campo de las broncoscopias virtuales y *Electromagnetic Navigation Bronchoscopy* (ENB), que utiliza campos electromagnéticos para la localización en tiempo real. La primera, es capaz de crear un entorno 3D capaz de replicar las vías aéreas con mucha precisión a través de una TC que favorece a la planificación y personalización de la intervención para cada paciente [21]. La segunda, la navegación electromagnética, es el máximo exponente en esta área. Desarrollada por el referente en la fabricación de equipamiento médico Medtronic bajo el nombre comercial de *SuperDimension*™ [22], además de realizar broncoscopias virtuales posee localizadores electromagnéticos que proporcionan la posición de la punta del broncoscopio en tiempo real. Existen varios estudios que afirman que estas técnicas aumentan el ratio de éxito en determinadas intervenciones, como, por ejemplo, la extracción de biopsias, aumentando el ratio diagnóstico y minimizando la tasa de neumotórax hasta un 1.5% frente a biopsias transbronquiales [22], [23], [24]. Sin embargo, estas técnicas presentan varias limitaciones importantes, como la necesidad de infraestructura adicional a la broncoscopia convencional, imágenes adicionales de tipo tomográfico que no deben ser de más de 2 semanas de antigüedad, así como equipamiento especializado costoso (mesa de trabajo especial, sensores y guías de broncoscopio especializadas) [22].

Todos estos métodos presentan ventajas y limitaciones. No hay un método óptimo o mejor que los demás, pues todos ellos se encuentran aún en desarrollo. Las limitaciones

de cada opción dificultan la realización de la técnica de broncoscopia de forma generalizada. Por ejemplo, los **procedimientos y equipamiento externos a la imagen** necesitan de un complejo procedimiento de planificación, impidiendo por ejemplo las intervenciones de urgencia y/o en tiempo real. Las herramientas de reconocimiento automático de imágenes resultan muy útiles en el campo de la medicina para ayudar en la identificación de estructuras y elementos tanto patológicos como no patológicos. Además, podrían proporcionar una vía alternativa para simplificar la navegación pulmonar en base únicamente a las imágenes de broncoscopia, sin la necesidad de equipamiento adicional ni planificación previa. Sin embargo, la **asistencia a la navegación por algoritmos de IA** todavía se encuentra en sus estadios más tempranos y su robustez todavía no es suficiente como para implementarlos en la práctica clínica. Finalmente, la **imagenología Híbrida y RAB**, que es lo más utilizado en la actualidad, tiene como limitaciones la peligrosidad relacionada con la radiación ionizante para los profesionales y toda la infraestructura robótica, que además de ser costosa es muy voluminosa. En la Figura 3 podemos apreciar toda la infraestructura requerida durante una broncoscopia asistida robóticamente.

Figura 3. Intervención broncoscópica asistida robóticamente realizada en el Hospital Germans Trias i Pujol, Badalona.

1.4 Aprendizaje profundo y sus aplicaciones

El aprendizaje profundo o *deep learning* (DL) es una rama del aprendizaje automático o *machine learning* (ML), siendo esta a su vez una parte de la inteligencia artificial (IA). Este tipo de aprendizaje se caracteriza por el uso de redes neuronales artificiales (RNA) de múltiples capas y gran profundidad, diseñadas para reconocer patrones y representaciones jerárquicas en grandes conjuntos de datos [25]. A diferencia de los algoritmos de aprendizaje más tradicionales, el DL automatiza la extracción de características relevantes, minimizando la intervención humana directa y capturando las relaciones entre los datos de estudio.

En el contexto de la IA y el aprendizaje profundo, existen numerosas tareas para las cuales hacer uso de RNA resulta especialmente útil. Aunque en sus orígenes estos algoritmos estaban orientados a realizar clasificaciones binarias o resolver problemas lógicos básicos [26], en la actualidad se aplican a una amplia gama de problemas y dominios:

- **Clasificación:** asignación de etiquetas a datos de entrada a la red.
- **Regresión:** predicción de valores continuos a partir de los datos de entrada.
- **Detección de objetos:** localización y clasificación de regiones de interés, principalmente de imágenes.
- **Reconocimiento de secuencias:** análisis y procesado de series temporales.
- **Procesamiento del lenguaje natural:** procesamiento, generación y traducción de textos.
- **Generación de datos:** creación de datos sintéticos a partir de datos reales.
- **Segmentación:** clasificación de píxeles en función del contexto espacial.

Si bien para cada tarea es necesario adaptar la arquitectura o el enfoque al problema de estudio, existen varias arquitecturas principales cuyo rendimiento las hace prevalecer sobre otras para optimizar distintas tareas. Las estrategias más utilizadas son:

1. **Redes Neuronales Convolucionales (CNN).** Este tipo de redes utiliza filtros convolucionales para detectar y reconocer patrones espaciales como bordes y texturas. Se utiliza ampliamente, especialmente para la clasificación y segmentación, con arquitecturas como la U-Net.
2. **Redes Neuronales Recurrentes (RNN).** Este tipo de redes se caracteriza por poseer “memoria”. Es decir, es capaz de almacenar información sobre estados

anteriores y utilizarla para procesar secuencias con contexto temporal. Una variante muy destacada de las RNN es la arquitectura LSTM (*Long Short-Term Memory*), que es capaz de mantener información relevante durante más tiempo.

3. **Transformers.** Este tipo de arquitecturas es de los más novedosos y utilizados en la actualidad, sobre todo para el procesamiento del lenguaje natural. Destaca por sus mecanismos de atención "selectiva" a los elementos más importantes de secuencias de texto, ponderando por ejemplo las palabras de una oración según su relevancia. También se utiliza en el área de la visión por computador, donde destacan los ViTs, o *Vision Transformers*.
4. **Autoencoders y modelos generativos.** Esta estrategia se utiliza mayormente para la creación de datos sintéticos o reconstrucción de imágenes. El mayor exponente en esta área son las redes de tipo GAN (*Generative Adversarial Networks*), donde se entrenan dos redes que compiten entre sí. Una red genera imágenes sintéticas y la otra las clasifica como reales o no, mejorando progresivamente la calidad y realismo de las imágenes generadas.

Dentro del marco de este TFG, y con la intención de realizar una segmentación automática de imágenes, es preciso ahondar en el concepto de las CNN, y sobre todo en la arquitectura de tipo U-Net. Como acabamos de comentar, las CNN permiten identificar bordes y texturas a partir de filtros convolucionales. Sin embargo, a medida que se pierde resolución durante este proceso, es complicado identificar y clasificar píxel por píxel el tipo de estructura al que pertenece la región seleccionada.

Para solventar este problema, Ronneberger *et al*. [27] propusieron la arquitectura **U-Net**. Esta arquitectura con forma de "U" cuenta con dos etapas, una de codificación, en la que se reduce la resolución y se van capturando las relaciones espaciales, y otra de decodificación, en la que se recupera la resolución original utilizando el contexto aprendido. Además, esta arquitectura permite el flujo de información de una etapa a otra mediante *skip connections*, lo que soluciona el problema de la pérdida de contexto durante la compresión de las imágenes y favorece la producción de segmentaciones más precisas.

1.5 Segmentación automática de imagen médica

La segmentación de imágenes es una tarea que consiste en identificar, delinear y separar estructuras o regiones de interés presentes en una imagen o volumen. Existen numerosos tipos y formas de segmentación. Sin embargo, para que este proceso se considere automático, se entiende que no debe existir intervención humana.

Para facilitar la segmentación, las regiones a delinear deben tener relevancia en el contexto de aplicación y no solaparse entre ellas. Dentro de las estrategias clásicas de segmentación automática encontramos algunas como la detección de bordes (a partir de filtros basados en la primera derivada), umbralización de intensidades de la imagen o el uso de funcionales de energía como *active contours* [28], [29], [30]. Estas estrategias, si bien son suficientemente precisas y útiles para algunas aplicaciones, muchas veces resultan insuficientes para la segmentación automática de imagen médica.

Es por esto por lo que se necesitan nuevas estrategias de segmentación que consigan resultados más precisos y escalables. De aquí surge la segmentación basada en *deep learning* o aprendizaje profundo. Por medio del procesamiento masivo de datos y la optimización de una función de pérdida, se pueden entrenar redes neuronales que al recibir como *input* una imagen den una salida en forma de segmentación. Para crear una red de esta índole se requieren grandes conjuntos de datos etiquetados por profesionales y un proceso iterativo de ajuste de hiperparámetros y entrenamiento.

Enlazado a la segmentación basada en *deep learning*, esta a su vez se subdivide en tres tipos de segmentación: la segmentación semántica, la segmentación por instancias y la segmentación panóptica [31], [32], [33].

- Segmentación semántica. Este tipo de segmentación consiste en clasificar y etiquetar cada píxel de una imagen o volumen en una clase distinta [31]. Esta estrategia es especialmente relevante en el contexto de la segmentación de imagen médica, puesto que es común intentar segmentar por ejemplo diferentes órganos a partir de una imagen de TC, donde cada píxel puede pertenecer a una estructura u otra. Para este tipo de segmentación, la arquitectura más relevante es la U-Net [27], que será explicada en la metodología de este trabajo.

- Segmentación por instancias. La segmentación por instancias resulta de la unión de clasificar cada píxel en una clase y, además, determinar el número de elementos o instancias que existen de esa misma clase en la imagen segmentada. Esta tarea es bastante más complicada que la anterior, y todavía supone un gran reto para la comunidad científica. Redes como *Mask Region-based Convolutional Neural Networks* (Mask R-CNN) son de gran ayuda para esta segmentación [33]. La segmentación por instancias se puede utilizar en el ámbito médico para realizar conteos de células o segmentar tumores de manera individual.
- Segmentación panóptica. La segmentación panóptica se define como la unión de los dos tipos de segmentación anteriores, aunando las mejores características de cada uno de ellos. En primer lugar, realiza una segmentación por instancias para detectar los elementos de interés y posteriormente los clasifica en una determinada clase de manera semántica. A continuación, divide la imagen entre regiones contables (elementos, e.g., tumores), y regiones amorfas (descrito como *stuff*), como por ejemplo tejido conectivo o fluido. Es un tipo de segmentación que pretende caracterizar a una imagen en su totalidad. Una de las arquitecturas más conocidas respecto a la segmentación panóptica es FPN (*Panoptic Feature Pyramid Network*) [33], [34].

Para el sistema respiratorio, la segmentación de imagen médica es mayoritariamente semántica. En la actualidad se han segmentado numerosas estructuras mediante *deep learning*, desde los pulmones hasta los vasos sanguíneos de estos. El estado del arte respecto a la segmentación de imágenes del sistema respiratorio se divide en dos partes. La primera se centra en la segmentación de patologías como nódulos pulmonares o derrames pleurales mediante de imágenes de RX (rayos X) o de TC. La segunda parte, segmenta las propias estructuras de este sistema, logrando rendimientos predictivos de segmentación de más del 90% para la mayoría de las estructuras bajo estudio. Existen numerosos trabajos que segmentan las vías aéreas, los vasos sanguíneos, los pulmones como conjunto y los pulmones divididos en lóbulos o segmentos; principalmente para imágenes de TC. Para todos ellos la estrategia utilizada deriva de las CNN, más específicamente de la U-Net o arquitecturas derivadas de esta última.

Sin embargo, pese a todas las investigaciones en el área de la segmentación del sistema respiratorio, todavía no se ha realizado una segmentación automática de imágenes

de broncoscopia, en parte debido a la falta de datos para entrenar arquitecturas que pudieran realizar esta tarea.

A día de hoy existen muchos *datasets* de segmentación para radiografías o imágenes de TC. Uno de los *datasets* más completos es Lung3D. Gracias a este *dataset* publicado por Kuang *et al.* y Xie *et al.* podemos visualizar la reconstrucción tridimensional de las distintas partes de la anatomía respiratoria en el visualizador *3DSlicer*. Cada estructura ha sido segmentada manualmente con la finalidad de crear un algoritmo para segmentar automáticamente los segmentos pulmonares por medio de funciones implícitas [35], [36].

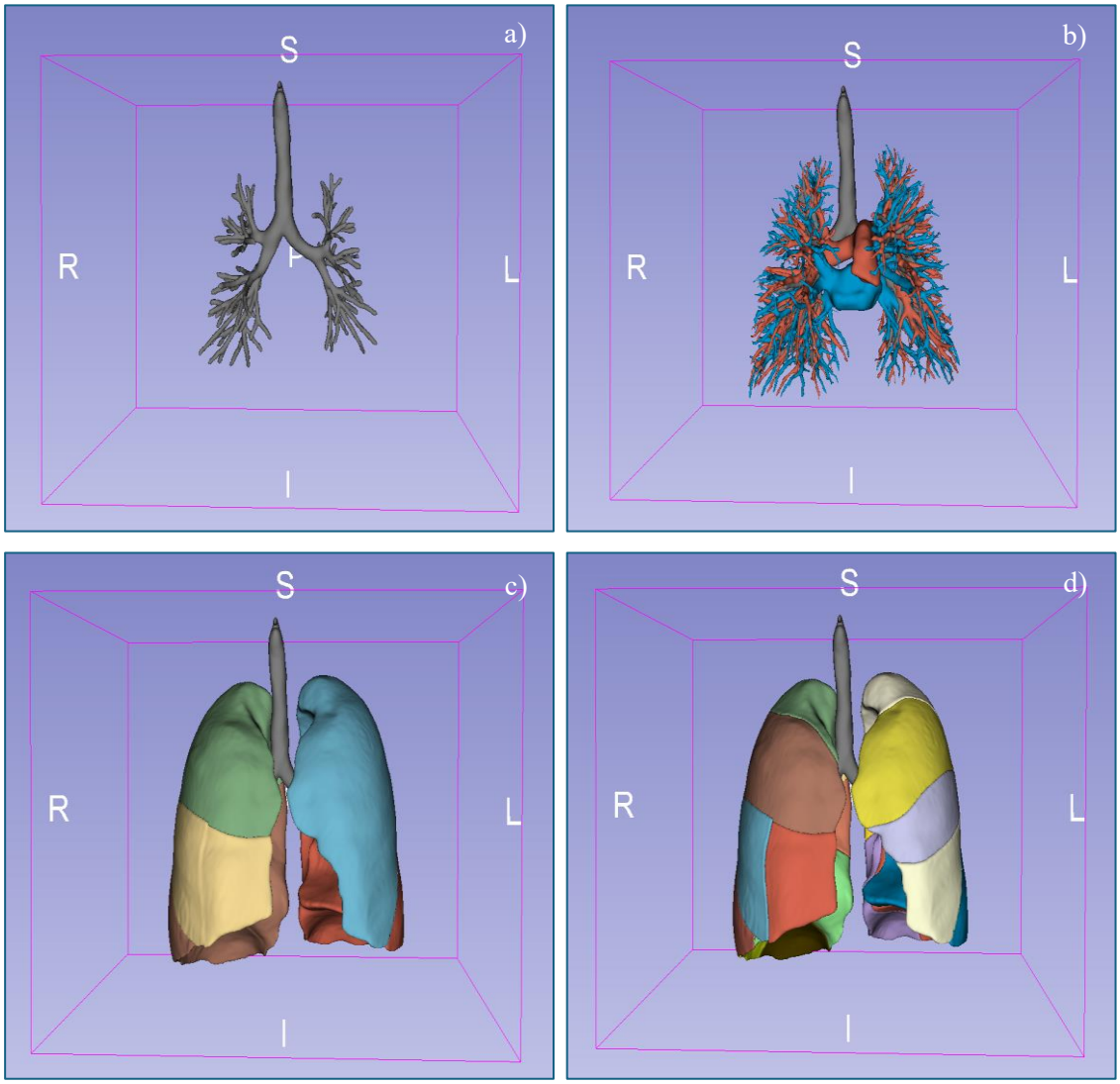

Figura 4. Reconstrucción 3D de: a) segmentación del árbol bronquial; b) árbol bronquial y su sistema circulatorio asociado; c) pulmones y su división en lóbulos; y d) pulmones y su división en segmentos asociados a los bronquios segmentarios. Dataset Lung3D [36].

En la Figura 4 (a) se puede apreciar toda la complejidad y ramificaciones del árbol bronquial. Del bronquio derecho se bifurcan los tres bronquios lobares y sus respectivos bronquios segmentarios. Del bronquio izquierdo se bifurcan los dos bronquios lobares y sus respectivos bronquios segmentarios. En las Figuras 2 (b), (c) y (d) se pueden apreciar las estructuras accesorias que completan toda la anatomía respiratoria.

2. Hipótesis y objetivos

Bajo el contexto del capítulo anterior surge la necesidad de conocer la localización exacta del broncoscopio en el árbol bronquial durante intervenciones endoscópicas del tracto respiratorio. No existe ningún método novedoso que domine sobre los demás de manera clara. Los mecanismos externos a la imagen o bien son muy costosos o requieren de demasiada infraestructura. Los algoritmos de asistencia a la navegación mediante IA se encuentran en pleno desarrollo y prometen ser el futuro, pero todavía no se usan en la práctica clínica por su falta de robustez. Finalmente, los métodos más utilizado son los de imagenología híbrida, pero la evidencia científica afirma que no son la mejor solución y que es necesario un enfoque multidisciplinar [16], [19], [37].

Por todo ello, en el presente Trabajo Fin de Grado (TFG) se propone una alternativa todavía no explorada en profundidad por la comunidad científica en relación a la broncoscopia: la segmentación semántica y automática de imágenes de broncoscopia para la localización bronquial. El procesado automático de imágenes de broncoscopia mediante técnicas de aprendizaje profundo podría ser una herramienta útil para la segmentación del árbol bronquial capaz de simplificar la localización del broncoscopio dentro de la estructura anatómica de la vía aérea.

2.1 Hipótesis

La segmentación automática de imagen médica en la actualidad se utiliza para segmentar estructuras anatómicas de imágenes o volúmenes CT o RM. En neumología se utiliza para segmentar los pulmones, las vías aéreas, venas y arterias, lóbulos y segmentos pulmonares. Numerosos estudios han obtenido muy buenos resultados en todas esta áreas [35], [38], [39], [40]. Sin embargo, todavía no se ha aplicado en imágenes de broncoscopia. Es por esto por lo que la hipótesis de partida ***(H1)*** se fundamente en que: *"las nuevas arquitecturas de aprendizaje profundo pueden ser capaces de identificar y segmentar las distintas estructuras anatómicas en una imagen de broncoscopia, así como proporcionar su localización dentro de la vía aérea a partir únicamente de imágenes segmentadas de distintas zonas del árbol bronquial".* Esto podría resultar de utilidad para identificar zonas de interés, ya sea para localizar lesiones, realizar biopsias y asistir de

forma general a los Neumólogos durante la broncoscopia, así como para entrenar a futuros especialistas en esta técnica [8].

2.2 Objetivos

2.2.1 Objetivo general

En base a esta hipótesis, se establece como ***objetivo general*** de este TFG **diseñar y validar un modelo predictivo para la detección y delineación automática de las estructuras naturales de la vía aérea** en imágenes capturadas durante la broncoscopia, para asistir al especialista durante el procedimiento. Para ello se implementarán y compararán diferentes arquitecturas de aprendizaje profundo para la segmentación semántica de las imágenes, en concreto dos arquitecturas de tipo U-Net: la U-Net convencional [27], [41] y la ResEnc U-Net, con conexiones residuales en el *encoder* [42].

2.2.2 Objetivos específicos

Para alcanzar el objetivo general, se proponen los siguientes objetivos específicos:

I. Evaluar y comparar algoritmos de DL para determinar cuál de las arquitecturas alcanza el mayor rendimiento predictivo en la delineación de las estructuras de la vía aérea.
II. Determinar qué estructuras anatómicas de la vía aérea superior y del árbol bronquial son identificadas con mayor precisión hasta la primera bifurcación. Este grupo de estructuras está conformado por 4 clases: (i) bronquio izquierdo; (ii) bronquio derecho; (iii) carina traqueal; (iv) background.
III. Localizar la sonda de broncoscopia dentro de la vía aérea a partir del procesamiento de las segmentaciones obtenidas mediante los algoritmos predictivos.

3. Materiales y diseño del estudio

En este capítulo se describen los materiales utilizados, así como la metodología empleada para el procedimiento de experimentación en referencia a la hipótesis planteada. Primeramente, se describe el diseño del experimento, así como la base de datos recopilada. Posteriormente se definen los aspectos éticos y la aprobación del comité para realizar el estudio. Finalmente se describe el proceso desde la grabación de los vídeos de broncoscopia hasta el momento de realizar la segmentación manual.

3.1 Diseño del estudio

En este TFG se ha realizado un estudio prospectivo observacional de creación y validación de modelos predictivos automáticos. Para ello se han recopilado los datos (vídeos de broncoscopia) de forma prospectiva por el neumólogo Milko D. Terranova en el Hospital Universitario Río Hortega (HURH) de Valladolid. La población objetivo del estudio está constituida en base a los siguientes criterios de inclusión y exclusión:

(i) Criterios de inclusión: pacientes mayores de 18 años que hayan aceptado participar en el estudio (firmando el consentimiento informado).

(ii) Criterios de exclusión: pacientes con intervenciones quirúrgicas previas que hayan afectado a la morfología de la vía aérea; pacientes que presenten variantes anatómicas muy poco comunes que pudiesen dificultar el aprendizaje automático; ser menor de edad y/o no firmar el consentimiento informado.

Por lo tanto, la población objetivo está compuesta por pacientes mayores de edad sin alteraciones relevantes en la anatomía de las vías aéreas que hayan aceptado participar. El periodo de reclutamiento de pacientes y recopilación de datos se fijó en 5 meses, entre el 1 de enero y el 31 de mayo de 2025, ajustado a los tiempos de realización del TFG. El protocolo se pudo poner en marcha de acuerdo a la práctica clínica de la unidad de técnicas del Servicio de Neumología condicionado por diversos factores (navidades, vacaciones del experto de técnica de broncoscopia, pacientes dentro de los criterios de inclusión, etc.), de forma que el primer vídeo se obtuvo a finales de enero.

Con el objetivo de caracterizar de manera breve la población de estudio, **para cada paciente se han recopilado las siguientes variables**: número de historia clínica, edad, sexo, fecha de la prueba y motivo de realización.

El vídeo de broncoscopia se centró en una exploración rápida de la anatomía del árbol bronquial. De forma protocolizada, en cada intervención el experto parte desde la tráquea hasta visualizar el bronquio derecho. Dentro de éste se visitan todas las bifurcaciones y se finaliza volviendo a la tráquea, realizando posteriormente la misma exploración por el bronquio izquierdo hasta volver de nuevo a la tráquea. Todo el procedimiento se realiza mientras la patología del paciente y el desarrollo normal de una broncoscopia lo permitan.

3.1.1 Equipamiento de broncoscopia

Para realizar el **procedimiento de broncoscopia**, el Servicio de Neumología del HURH dispone del siguiente equipamiento:

- **Equipo OLYMPUS** (consola ecográfica, unidad de transmisión y sondas) para la realización de **broncoscopia convencional y ecobroncoscopia (EBUS)**, que incluye: Procesador Olympus con fuente de luz Evis Exera III CV-190 PLUs y Evis Exera III CLV-190 con conector estanco de un movimiento y teclado (para broncoscopia convencional); sistema de ecoendoscopia universal Olympus EU-ME-2 Premier con teclado (para EBUS lineal y radial); vídeo-broncoscopio pediátrico Olympus BF TYPE P 180 (4.9 mm), con canal de trabajo de 2.0 mm; vídeo-broncoscopio Olympus BF Q 190 (4.9 mm), con canal de trabajo de 2.0 mm; vídeo- broncoscopio terapéutico Olympus BF 1T180 (6 mm), con canal de trabajo ancho 3.0 mm; vídeo-broncoscopio ultrafino Olympus MP190 BF ultrafino 3.0 mm, con canal de trabajo 1.7 mm; ecobroncoscopio BF Type UC180F (6.3 mm y 6.9 mm de punta distal) con canal de trabajo 2.2 mm para EBUS lineal; unidad de transmisión (Motor) y minisonda radial 1.7mm para EBUS radial.
- **Equipo AMBU** (broncoscopios desechables y pantalla) para la realización solo de **broncoscopia convencional**, que incluye: vídeo-broncoscopios desechables AMBU®aScope 5; monitor-pantalla Ambu® aView 2 Advance; broncoscopios

AMBU aScope 5, terapéutico 5.6 mm (canal de trabajo 2.8 mm), fino 4.2 mm (canal de trabajo 2.2 mm) y ultrafino 2.7 mm (canal de trabajo 1.2 mm).

Ambos equipos permiten obtener y descargar en diferentes formatos tanto vídeos (Olympus: AVI, MPG; AMBU: MP4) como imágenes (Olympus: JPG, TIFF; AMBU: PNG). Los vídeos grabados tanto por Olympus como por AMBU están disponibles con un *framerate* de 30 fotogramas por segundo.

3.2 Aspectos éticos

Toda la metodología y aspectos éticos de este proyecto han sido evaluados y aprobados por el Comité de Ética de la Investigación con medicamentos de las Áreas de Salud de Valladolid, emitiendo un dictamen favorable el 12 de febrero de 2025 (acta nº 2 de 2025). El código asociado a este proyecto es PI-25-76-H.

El reclutamiento prospectivo de los pacientes se ha realizado de acuerdo con la Declaración de Helsinki sobre principios éticos para la investigación médica con sujetos humanos, así como con la resolución del Consejo de Europa sobre derechos humanos y biomedicina (CETS Nº 195, 2005). No se ha realizado ninguna intervención al paciente que no forme parte de la práctica clínica habitual y se ha obtenido el consentimiento informado firmado por todos los pacientes del estudio o los responsables de estos.

La investigación propuesta se ajusta a la legislación vigente (Ley 14/2007, de Investigación Biomédica y Real Decreto 2132/2004) del Gobierno español, así como a los principios éticos de la Carta de Derechos Fundamentales de la Unión Europea (2000/C 364/01) y los del Grupo Europeo de Ética de la Ciencia y las Nuevas Tecnologías (SEC(97)2404). Toda la información (datos clínicos e imágenes) ha sido tratada de acuerdo con el Reglamento General de Protección de Datos (RGPD) de la Comisión Europea (UE 2016/679) y la Ley Orgánica de Protección de Datos Personales y Garantía de los Derechos Digitales (LOPDGDD) de España (3/2018).

Toda la información de los pacientes (datos clínicos e imágenes) ha sido anonimizada durante la adquisición y los investigadores no han tenido acceso a los datos personales de los mismos.

3.3 Base de datos de broncoscopias

La base de datos final para este trabajo está formada por todo el conjunto de broncoscopias recopiladas en el HURH a raíz del protocolo descrito anteriormente. Se recopilaron un total de 35 vídeos, de los que 2 se han obtenido con el broncoscopio desechable AMBU y los 33 restantes con el broncoscopio OLYMPUS. Si bien ambos broncoscopios permiten evaluar las vías aéreas de la misma forma, existen diferencias en la visualización debido a los ángulos de grabación de los endoscopios. En la Figuras 5 podemos apreciar estas diferencias. Aunque ambas imágenes muestran la primera bifurcación a nivel de la carina traqueal, se puede apreciar cómo la imagen tomada con el OLYMPUS es más cálida y tiene mayor luminosidad. Además, este broncoscopio proporciona mayor resolución. Mientras que los vídeos tomados con el broncoscopio AMBU tienen una resolución de 400 × 400 píxeles, los vídeos tomados con el OLYMPUS poseen una resolución de alta definición, siendo de 1920 × 1080 píxeles. Otro aspecto a tener en cuenta es que, mientras que con el equipamiento AMBU el vídeo comprende únicamente la imagen de la broncoscopia, con el equipamiento OLYMPUS aparecen además en la pantalla datos como la fecha y hora, así como una imagen superpuesta sobre la propia broncoscopia, dificultando el análisis.

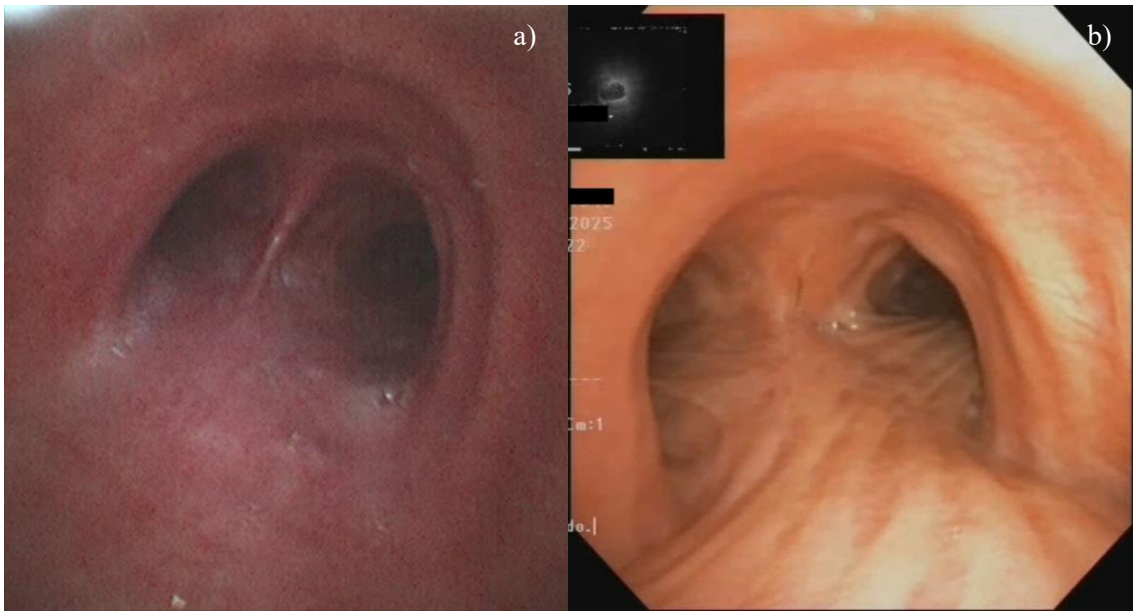

Figura 5. Comparación de las imágenes de broncoscopia: (a) imagen con equipamiento AMBU; (b) imagen con equipamiento OLYMPUS.

3.3.1 Población de estudio

La población de estudio está compuesta por 35 pacientes, cada uno contribuyendo con un vídeo de broncoscopia. El proceso de grabación se ha centrado en visualizar toda la anatomía del paciente en el menor tiempo posible para no alterar la realización normal de la prueba. Es por esto por lo que, en algunos pacientes, debido a su patología, los vídeos han sido más cortos o de menor calidad, con la intención de reducir al máximo el tiempo de la broncoscopia. Estos vídeos no se consideran casos incompletos, sino que simplemente no permiten visualizar toda la anatomía.

Respecto a la duración, los vídeos han tenido una duración mínima de 55 segundos y una duración máxima de 3 minutos y 6 segundos. La duración media ha sido de 84.8 ± 28.88 segundos.

Respecto a la población, la edad media de los pacientes fue de 59.61 ± 14.15 años. El paciente más joven presentó una edad de 29 años y el paciente de mayor edad 84 años. De los 35 pacientes, 19 fueron mujeres, lo que supone un 54.29% de la muestra, prácticamente balanceada con respecto al sexo. Los 16 hombres restantes representaron un 45.71% de nuestra población. De todos los pacientes, el 68.57% (24 pacientes) presentaron variantes anatómicas. La variante más común fue la trifurcación culmen del Lóbulo Superior Izquierdo LB1, LB2 y LB3, cuando lo normal sería una bifurcación de la que parten LB1 + LB2 por un lado y LB3 por el otro.

Las indicaciones para la broncoscopia han sido variadas. En la Tabla 2 se muestra una recopilación de los principales motivos de realización, así como sus respectivos porcentajes dentro del conjunto poblacional. Se puede apreciar cómo las patologías más comunes que terminan en broncoscopia dentro de nuestra población muestral son las enfermedades infiltrativas pulmonares, con una representación del 25.7%, es decir, aquellas que afectan al intersticio; así como las que afectan directamente a las vías aéreas, principalmente por la inflamación de estas, con un 20% del total.

Tabla 2. Clasificación de las enfermedades más recurrentes en la población de estudio

	Número de observaciones	*% del total de pacientes*	*% del total de tipos de consulta*
Tipo de patología			
Relación con hematología	**5**	**14,3**	
Hematemesis	1	2,9	20,0
Expectoración hemoptoica	1	2,9	20,0
Hemoptisis	3	8,6	60,0
Infiltrados	**9**	**25,7**	
EPID	3	8,6	33,3
Neumonía	4	11,4	44,4
Patrón Intersticial Sarcoideo	1	2,9	11,1
Infiltrados sin especificar	1	2,9	11,1
Infección y patrones febriles	**6**	**17,1**	
Infección respiratoria	4	11,4	66,7
Síndrome febril	1	2,9	16,7
Picos de fiebre	1	2,9	16,7
Relación con Neoplasia	**5**	**14,3**	
Nódulo pulmonar	3	8,6	60,0
Sospecha de neoplasia	1	2,9	20,0
Síndrome constitucional	1	2,9	20,0
Inflamación de vías aéreas y derivados	**7**	**20,0**	
Asma	2	5,7	28,6
Hiperreactividad bronquial	2	5,7	28,6
Bronquiectasia	1	2,9	14,3
Tos	2	5,7	28,6
Colapso	**1**	**2,9**	
Atelectasia	1	2,9	100,0
No relacionado con vías aéreas propias	**1**	**2,9**	
Dolor dorsal	1	2,9	100,0
Tipo de consulta			
RVA	13	37,1	
Otros	22	62,9	

4. Metodología

En este capítulo se resume toda la metodología, algoritmos y técnicas empleadas en el flujo de trabajo, desde la obtención de las imágenes y el entrenamiento de los modelos hasta las métricas de rendimiento empleadas para analizar la calidad de la segmentación automática. En primer lugar, se describe el proceso de caracterización de los vídeos en matrices y su segmentación manual para crear las etiquetas de entrada a los distintos modelos de segmentación, además de la propia herramienta para realizar esta tarea. Una vez completado el *dataset* con las imágenes y sus etiquetas, así como su partición en entrenamiento - test, se introduce el concepto de U-Net en el aprendizaje profundo y el *framework* nn-Unet (no-new U-Net), de gran utilidad para las tareas de segmentación automática, con preprocesado, entrenamiento y postprocesado automatizados. Posteriormente se definen las arquitecturas de red que se utilizarán para la tarea de segmentación, los hiperparámetros, la función de pérdida y el procedimiento de *data augmentation*. Finalmente, se describe el postprocesado realizado, así como las métricas de rendimiento empleadas, hasta dar paso a la funcionalidad final de la segmentación automática: la localización dentro del árbol bronquial.

4.1 Creación del *dataset*

Un *dataset* o conjunto de datos es una recopilación de todos los datos extraídos y caracterizados para una tarea determinada. Como se describe a lo largo del trabajo, la tarea a realizar es la segmentación de las estructuras de la vía aéreas mediante técnicas de aprendizaje profundo. Con un enfoque de aprendizaje supervisado, el *dataset* está compuesto por los datos y sus etiquetas. Para este estudio, los datos son las imágenes extraídas y procesadas a partir de los vídeos, mientras que las etiquetas son la segmentación manual realizada sobre cada imagen delimitando cada estructura anatómica en el campo de visión de cada imagen.

4.1.1 Preprocesado de los vídeos

Una vez recopilados los vídeos, es necesario aplicar una serie de modificaciones previas al preprocesado para facilitar el trabajo posterior de creación del *dataset*. Gracias

a estos cambios será posible procesar los vídeos de una forma más rápida y sencilla, optimizando el espacio en memoria que ocupa cada vídeo y la rapidez de computación. La conversión de los vídeos originales ha consistido en: *(i)* transformación del vídeo RGB a matriz, *(ii)* recorte de las matrices y *(iii)* almacenamiento en formato *.nii.gz*.

(i) **Transformación del vídeo RGB a matriz.** Este paso consiste en la transformación del vídeo de broncoscopia RGB a matriz numérica de un único canal para trabajar con ella. Esto se ha realizado *frame* a *frame*, por lo que las imágenes no mantienen un contexto temporal entre ellas, siendo analizadas como imágenes por separado. Los vídeos RGB tienen una estructura *height*×*width*×*depth*×*channels*, donde existen 3 canales, uno para cada color (rojo, verde y azul) con valores de 0 a 255. Al transformar el video de RGB a un único canal, lo estamos llevando a la escala de grises, también de 0 a 255. Esto lo podemos conseguir eliminando la dimensión de los canales, transformando de *height*×*width*×*depth*×*3* a *height*×*width*×*depth*×*1*. El número 1 hace referencia a que existe un único canal, que en este caso nos dará una imagen en escala de grises. Si bien es cierto que estamos perdiendo información, esta transformación es común y ha sido contrastada en estudios de detección y clasificación de pólipos en imágenes de endoscopia, que demuestran que la perdida de información no es tan relevante, e incluso puede aumentar métricas como la precisión [43].

Además de esto, se han permutado las dimensiones de las matrices para tener la forma *n_images*×*height*×*width*, de acuerdo a la convención más utilizada normalmente en imágenes de tipo TC, donde *n_images* hace referencia al eje axial o *z*. Esta dimensión la hemos denominado como *n_images* debido a la falta de contexto temporal en el análisis. El resultado de este paso es una matriz de imágenes apiladas a lo largo de la primera dimensión, como si de una imagen TC se tratase, pero con imágenes independientes entre sí.

(ii) **Recorte de las matrices y redimensionado –** Como se ha mencionado con anterioridad, existen dos tipos de imágenes recopiladas, las del equipo AMBU y las del equipo OLYMPUS. Las imágenes del primer equipo no necesitan ser recortadas. Sin embargo, las de OLYMPUS requieren un recorte tanto en altura como en anchura, puesto que los vídeos contienen más información a mayores del propio vídeo. Tras el recorte, las imágenes pasan de tener 1920×1080

píxeles a tener 540×540. Con esto conseguimos imágenes cuadradas que contienen únicamente información relevante. Las imágenes del equipo AMBU han sido redimensionadas a 540×540 para tener todo el conjunto de imágenes con las mismas dimensiones. Las imágenes han sido redimensionadas mediante interpolación bi-cúbica y las etiquetas han sido redimensionadas mediante el algoritmo de interpolación *nearest-neighbour* [44].

(iii) **Almacenamiento en formato *.nii.gz* –** *Neuroimaging Informatics Technology Initiative* (NIfTI, con extensión *.nii*) es uno de los formatos más comunes de almacenamiento de imagen médica. Además, permite la compresión gzip. Gracias a esto es posible almacenar los vídeos en matrices de forma eficiente a nivel de almacenamiento y lectura.

Tras la adecuación de los vídeos, el resultado son matrices listas para la segmentación. En la Figura 6 podemos ver una comparación entre una captura del vídeo original (OLYMPUS) y una imagen ya preprocesada y en el formato final.

4.1.2 Segmentación manual y partición del *dataset*

Una segmentación manual correcta es indispensable para establecer un *ground truth* (GT) de calidad que dé lugar a predicciones sin errores. Ésta se realizará sobre las matrices caracterizadas y procesadas ya en formato *.nii.gz*. La segmentación para esta tarea es multiclase, ya que hay que segmentar varias estructuras con un mismo modelo. Además, no debe tener solapamiento, pues para este problema se sobreentiende que no hay estructuras superpuestas a otras. Es por esto por lo que la segmentación se almacenará por imágenes y no por etiquetas, es decir, una imagen podrá contener más de una etiqueta simultáneamente, y no una etiqueta por canal. De esta forma se consigue ahorrar en espacio de almacenamiento y se mejora la rapidez de procesamiento.

En la Tabla 3 se muestra el conjunto de etiquetas posibles a largo de todo un vídeo de exploración. Se puede apreciar cómo existen estructuras que se segmentan en conjunto con otras. Esto sucede con LB1 y LB2 y con la língula, que comprende LB4 y LB5. Además, para hacer segmentaciones fluidas y no tener imágenes sin segmentar se han creado las estructuras auxiliares RML + RLL y RB_RLL-78910.

La segmentación se ha realizado empleando la herramienta software de visualización y procesado de imágenes médicas en código abierto ITK-SNAP (http://www.itksnap.org). Esta herramienta ha sido seleccionada por su elevada precisión [45], [46] , así como por estar disponible de forma gratuita y ser *open source*. Con este software es posible importar y exportar el conjunto de etiquetas y sus valores, facilitando la selección de etiquetas entre distintos pacientes.

Para obtener el *dataset* final se han segmentado los 35 vídeos o pacientes. Sin embargo, debido a limitaciones en el tiempo disponible para realizar el trabajo, se ha segmentado en su totalidad únicamente los dos primeros pacientes. El resto de los pacientes tienen segmentado el bronquio izquierdo (LB), el bronquio derecho (RB) y la carina traqueal (CTraq). En la Figura 7 se muestran algunos ejemplos de imágenes segmentadas de todo el árbol bronquial.

Finalmente se ha segmentado un total de 8591 imágenes útiles para el entrenamiento y validación de la red, con las etiquetas RB, LB y Ctraq. Debido al carácter prospectivo de la recopilación de pruebas de broncoscopia se fijó un umbral temporal para obtener el dataset de entrenamiento y poder comenzar con el diseño y proceso de aprendizaje y optimización de los modelos, mientras que el dataset de test se fue completando hasta la fecha límite propuesta para la etapa de recopilación de datos. La distribución final ha sido de 19 pacientes para el entrenamiento (4787 imágenes) y 16 pacientes para el test (4164 imágenes), es decir, una proporción del 55% para entrenamiento y 45% para test.

4.2 Segmentación automática

Para segmentar automáticamente una imagen existen numerosos enfoques. Desde un enfoque de aprendizaje profundo, debemos crear una red neuronal convolucional (CNN). Una vez preprocesados los datos, debemos crear y entrenar la red con ellos y evaluar las predicciones de la red. Gracias a las U-Nets auto configurables y al *framework* Nn-Unet es posible automatizar gran parte de este proceso [41].

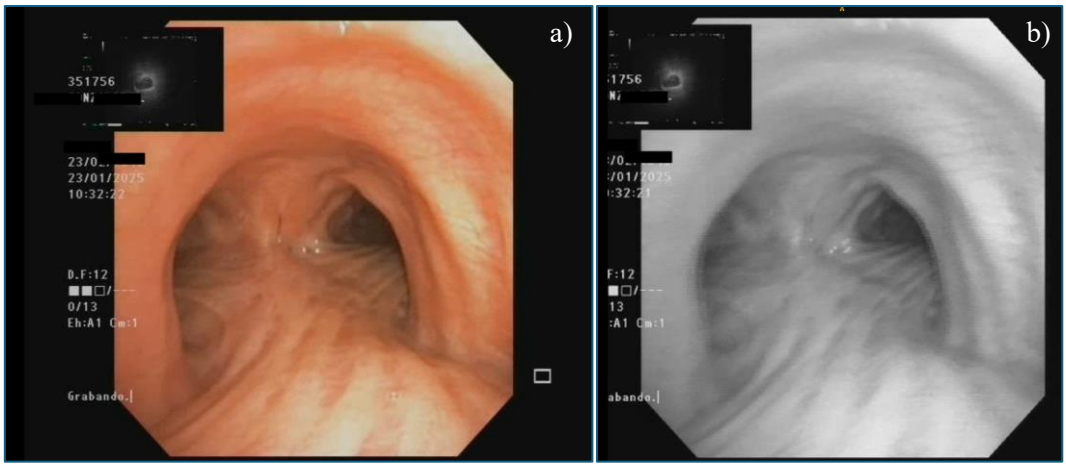

Figura 6. Comparación entre (a) captura del vídeo original y (b) imagen caracterizada y preprocesada.

Tabla 3. Etiquetas y valores asociados en la segmentación manual.

Etiqueta	**Valor**	**Etiqueta**	**Valor**
Clear Label	0	LB1 + LB2	14
LB	1	LB3	15
RB	2	Lingula (LB4 + LB5)	16
Ctraq	3	LB6	17
RB1	4	LB7	18
RB2	5	LB8	19
RB3	6	LB9	20
RB4	7	LB10	21
RB5	8	RUL	22
RB6	9	RML	23
RB7	10	RLL	24
RB8	11	LUL	25
RB9	12	LLL	26
RB10	13	RML + RLL	27
		RB – RLL - 78910	28

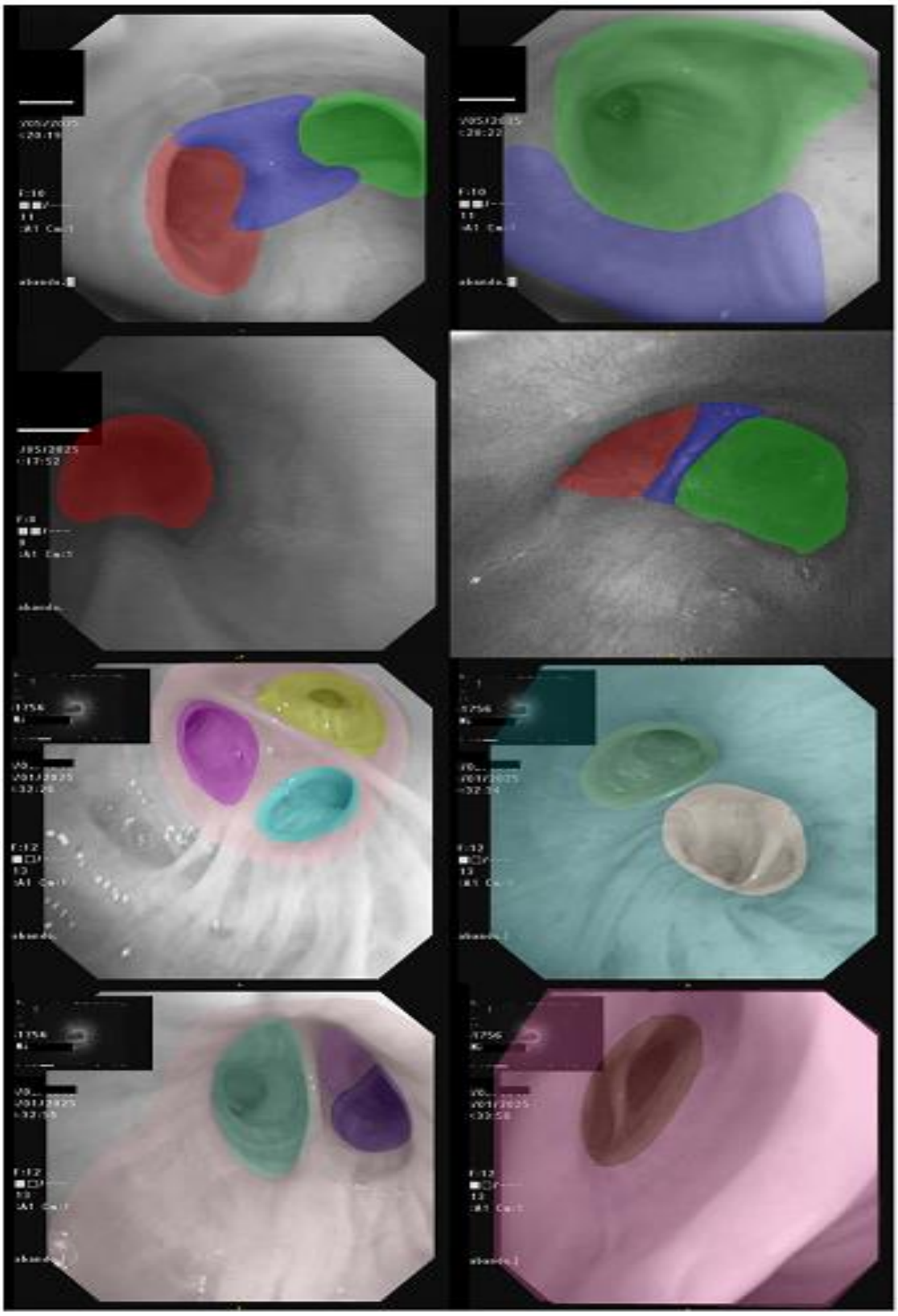

Figura 7. Ejemplos de estructuras segmentadas a lo largo de toda la exploración. De arriba a abajo, las cuatro primeras imágenes muestran Ctraq, LB y RB del equipamiento OLYMPUS y AMBU. Las cuatro últimas muestran i) RUL, RB1, RB2, RB3; ii) RML+RLL, RML, RLL; iii) RB8, RB9, RB10, RB_RLL_78910 iv) LII, LB6.

4.2.1 Aprendizaje profundo: Redes Neuronales Convolucionales

El aprendizaje profundo o *Deep Learning (DL)* es una estrategia de aprendizaje computacional relativamente reciente que se ha convertido en el paradigma estándar de procesado de imagen en general y, más específicamente, de la imagen médica. Este tipo de aprendizaje se caracteriza por una intervención humana limitada únicamente a la "preparación" de la información de entrada a los modelos, eliminando la etapa de *feature engineering*, que se realiza de forma automática en las primeras etapas. Las arquitecturas de *DL* se basan originalmente en la idea de perceptrón o neurona artificial, que, tratando de simular al funcionamiento biológico del cerebro humano y las neuronas, se utilizaba para resolver problemas de clasificación y separación lineal con dos únicas clases, como la lógica booleana. Fue introducido por McCulloch y Pitts en 1943 y perfeccionado por Rosenblatt en 1958, quien crea la primera asociación conceptual del aprendizaje supervisado con el perceptrón [26], [47].

Un perceptrón simple consiste en una única neurona (unidad de computación) con un peso y sesgo ajustables. Gracias a un algoritmo que ajusta los parámetros de la neurona mediante un proceso iterativo, se puede lograr una convergencia, permitiendo implementar tareas de clasificación.

Cuando se emplean varias neuronas para un mismo problema, suceden varias cosas: aumenta la complejidad de la arquitectura, pues cuantas más neuronas más no linealidad se introduce al clasificador; y más poder de computación y tiempo se requiere para la convergencia. A esta unión de neuronas se le denomina MLP (*Multilayer Perceptron)* o perceptrón multicapa. Con esta unión de neuronas se introduce el concepto de *backpropagation,* para entrenar eficientemente estas nuevas arquitecturas. Este proceso de aprendizaje ajusta iterativamente los pesos de las conexiones en la red, de forma que se minimice la diferencia entre el *output* actual de la red o predicción y la etiqueta u *output* esperado. En la Figura 8 se muestra una representación de un perceptrón multicapa [48]. En la imagen se aprecia el diagrama con las operaciones en forma de pasos y un modelo simplificado. En el diagrama (a) se representan *n* entradas con sus respectivos pesos, el sumatorio de todos ellos y su paso por la función de activación para dar lugar a una salida *y*.

Las arquitecturas de *deep learning* se basan en el perceptrón multicapa, pero están compuestas por muchas más capas, en las que se van identificando de forma automática las propiedades o patrones inherentes de las entradas (en el caso de este trabajo, imágenes). Esto se realiza a diferentes niveles de abstracción, sin que sea necesaria la intervención de expertos en el área correspondiente que tengan que parametrizar las propiedades de las imágenes mediante características (conjunto de variables independientes) basadas en el conocimiento del problema bajo estudio, en contraste con el enfoque de *feature engineering* convencional [47].

En relación al deep learning existen numerosas arquitecturas o estrategias. Para imágenes y volúmenes, la arquitectura CNN (Convolutional Neural Network) es la más ampliamente utilizada. Este tipo de arquitectura tradicionalmente permite extraer características espaciales y jerárquicas de las imágenes, como bordes, texturas y formas. Sin embargo, debido a su estructura en forma de embudo va reduciendo progresivamente la resolución espacial por medio de convoluciones. Las capas iniciales capturan características locales, mientras que las capas finales combinan estas características para tomar decisiones globales. Este tipo de redes puede perder precisión a nivel de información de detalle, aunque en la actualidad son utilizadas con gran rendimiento predictivo en problemas de clasificación y detección de objetos [49], [50]. En la Figura 9 se observa cómo afectan las convoluciones a la resolución espacial en las CNN para una arquitectura VGG (Visual Geometry Group), [51], [52]. Esta arquitectura es ampliamente utilizada en la comunidad científica, siendo una de las bases de la investigación para este tipo de redes.

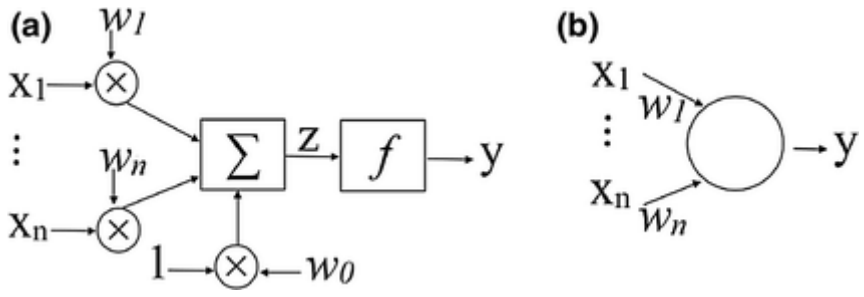

Figura 8. Modelo del Perceptrón, de derecha a izquierda: (a) modelo por pasos; (b) modelo simplificado. Adaptado de [39]

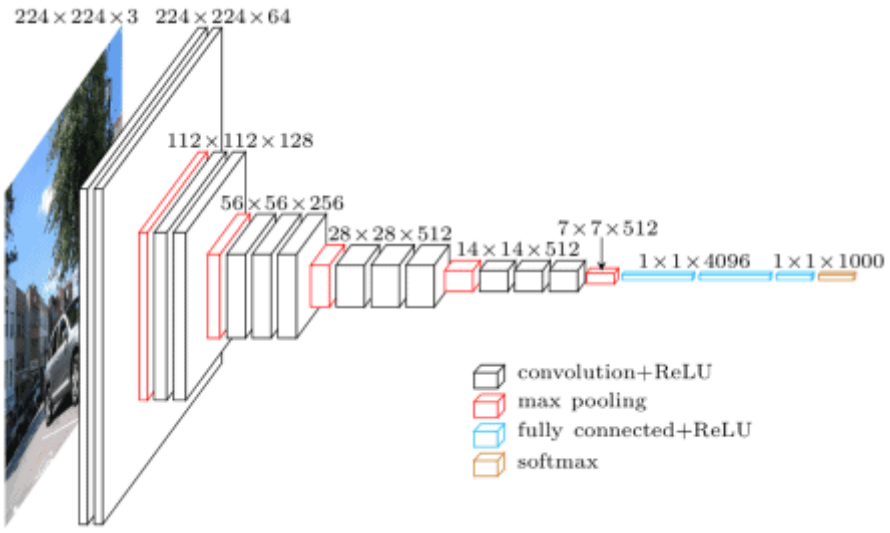

Figura 9. Estructura de VGG, representando una CNN. Extraído de [51]

4.2.1.1 Arquitectura U-Net

Para la tarea de segmentación de imagen existe en la literatura una arquitectura especial derivada de las CNN convencionales que se ha convertido en estándar de referencia en muy poco tiempo. Esta arquitectura, llamada U-Net, fue introducida por primera vez en 2015 por Ronneberger *et al* [27]. Esta arquitectura con forma de "U" tiene una primera parte denominada *encoder* o codificador, de forma similar a una CNN, en la que se van reduciendo las dimensiones espaciales y extrayendo características de la imagen; y una segunda parte llamada *decoder* o decodificador en la que se van recuperando las dimensiones espaciales originales a la vez que se combina con características de alta resolución procedentes de la etapa anterior. Esta combinación de características directamente desde el *encoder*, denominada *skip-connections,* es lo que permite combinar información de bajo nivel (detalles espaciales) con características de alto nivel (más abstractas) manteniendo la resolución espacial preservando detalles finos. Esto es ideal para identificar tejidos y estructuras en imagenología médica [53]. En la Figura 10 se muestra la arquitectura U-Net original propuesta por Ronnerberg *et al.* [27].

Si bien existen numerosas variaciones de esta arquitectura para la tarea de segmentación, diez años más tarde sigue siendo una de las más utilizadas por la comunidad, científica respaldada por sus buenos resultados.

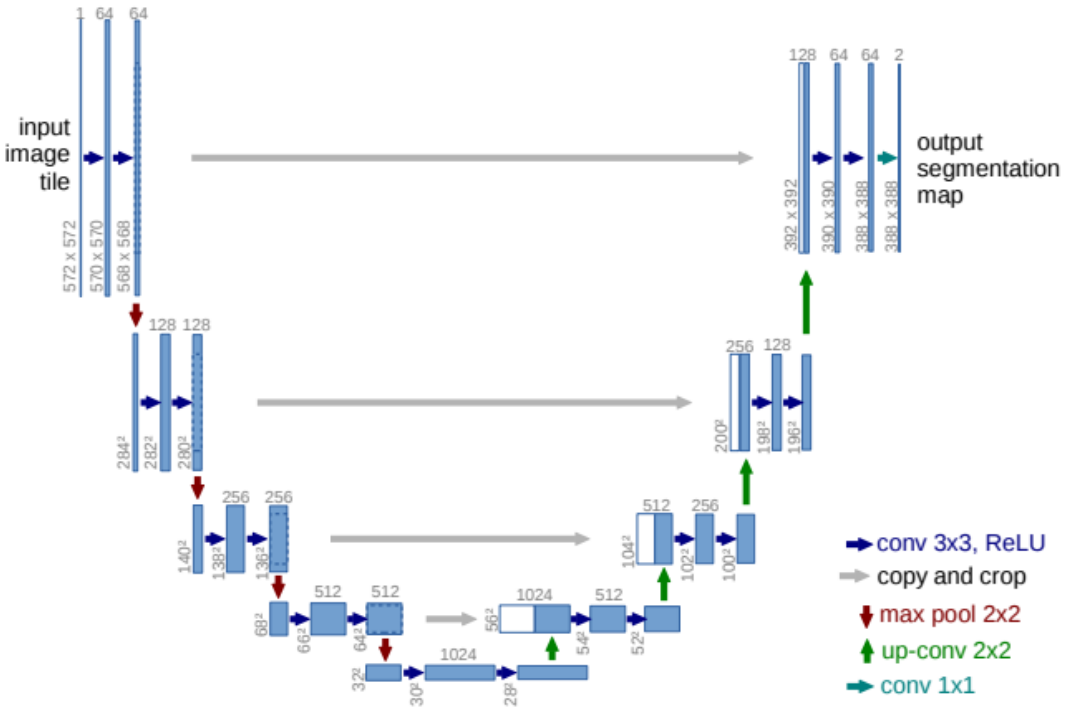

Figura 10. Arquitectura U-Net original o plain U-Net [31]. Cada caja azul corresponde a un mapa de características multicanal. Las cajas blancas representan la copia de los mapas de características de las skip-connections. Las flechas denotan el tipo de operación. Imagen extraída de [27].

4.2.1.2 Arquitectura ResEnc U-Net

En contraposición a la Plain U-Net o U-Net convencional, la arquitectura ResEnc U-Net contiene bloques residuales en el codificador en vez de una secuencia simple de convoluciones [54].

Los bloques residuales, introducidos por primera vez por Kaiming *et al.* en 2015, son uniones o conexiones directas del codificador a distintos niveles, que permiten "saltarse" una o más capas de la red, aprendiendo únicamente la diferencia entre la entrada y la salida deseadas, en lugar de tener que aprender toda la información desde cero. Resulta de especial utilidad en redes muy profundas para mejorar el problema del desvanecimiento de gradiente (actualizaciones del modelo demasiado pequeñas) [55]. En la Figura 11, se muestra un esquema del aprendizaje residual, donde existe una conexión entre distintas capas. Con ella, el gradiente se salta capas y la red es capaz de aprender más fácilmente relaciones entre patrones para redes muy extensas.

La arquitectura ResEnc U-Net aprovecha estas conexiones residuales para ofrecer redes de segmentación más densas y con más capas, mejorando los resultados ligeramente

respecto a la Plain U-Net. Ambas arquitecturas se comparan a nivel técnico en la Tabla 4. La Figura 12 muestra estas diferencias de manera gráfica. Una de las modificaciones más notables se produce en el bloque *encoder* o codificador, en el que la Plain U-Net repite una secuencia "convolución, normalización y función de activación" dos veces por bloque, mientras que ResEnc U-Net conecta el inicio de cada bloque con el final de manera residual hasta justo antes de la segunda función de activación. Además, Plain U-Net inicializa la red con 30 filtros, mientras que ResEnc U-Net inicializa en 24, para ahorrar algo de memoria. La principal ventaja que ofrecen las conexiones residuales frente a la U-Net original es una mejor propagación del gradiente a costa de un mayor consumo de memoria. Sin embargo, no demuestra mejoras demasiado significativas en el rendimiento predictivo según sus creadores [54]. Finalmente, en relación a los parámetros calculados en base a reglas, debido al mayor consumo de memoria por parte de la arquitectura, el *patch size* y el *batch size* se ajustarán automáticamente para mejorar el rendimiento de acuerdo al flujo de trabajo descrito a continuación.

Tabla 4.Comparativa entre Plain U-Net y ResEnc U-Net.

	Plain U-Net	**ResEnc U-Net**
Arquitectura base	U-Net convencional	U-Net + bloques residuales en encoder
Bloques *Encoder*	2 × Conv → InstNorm → ReLU	Conv → InstNorm → ReLU → Conv → InstNorm → +Residual → ReLU
Bloques *Decoder*	2 × Conv → InstNorm → ReLU	Conv → InstNorm → ReLU
Filtros en inicialización	30	24
Downsampling	Conv 3×3 (stride 2)	Conv 3×3 (stride 2)
Upsampling	Transposed Conv	Transposed Conv
Tipo de conexiones	Secuencial	Secuencial y residual
Uso de memoria	Relativamente bajo	Alto
Ventajas	Simple y eficiente	Mejora de propagación del gradiente
Desventajas	Estancamiento en redes profundas	Alto consumo de memoria

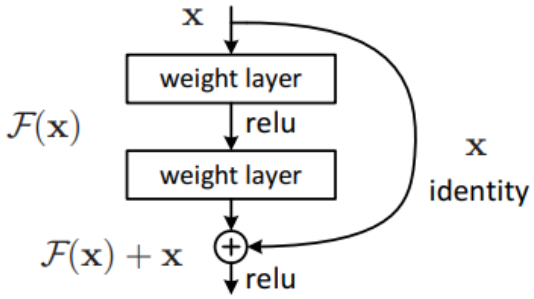

Figura 11. Aprendizaje residual: bloque de construcción de la red. Extraído de [55]

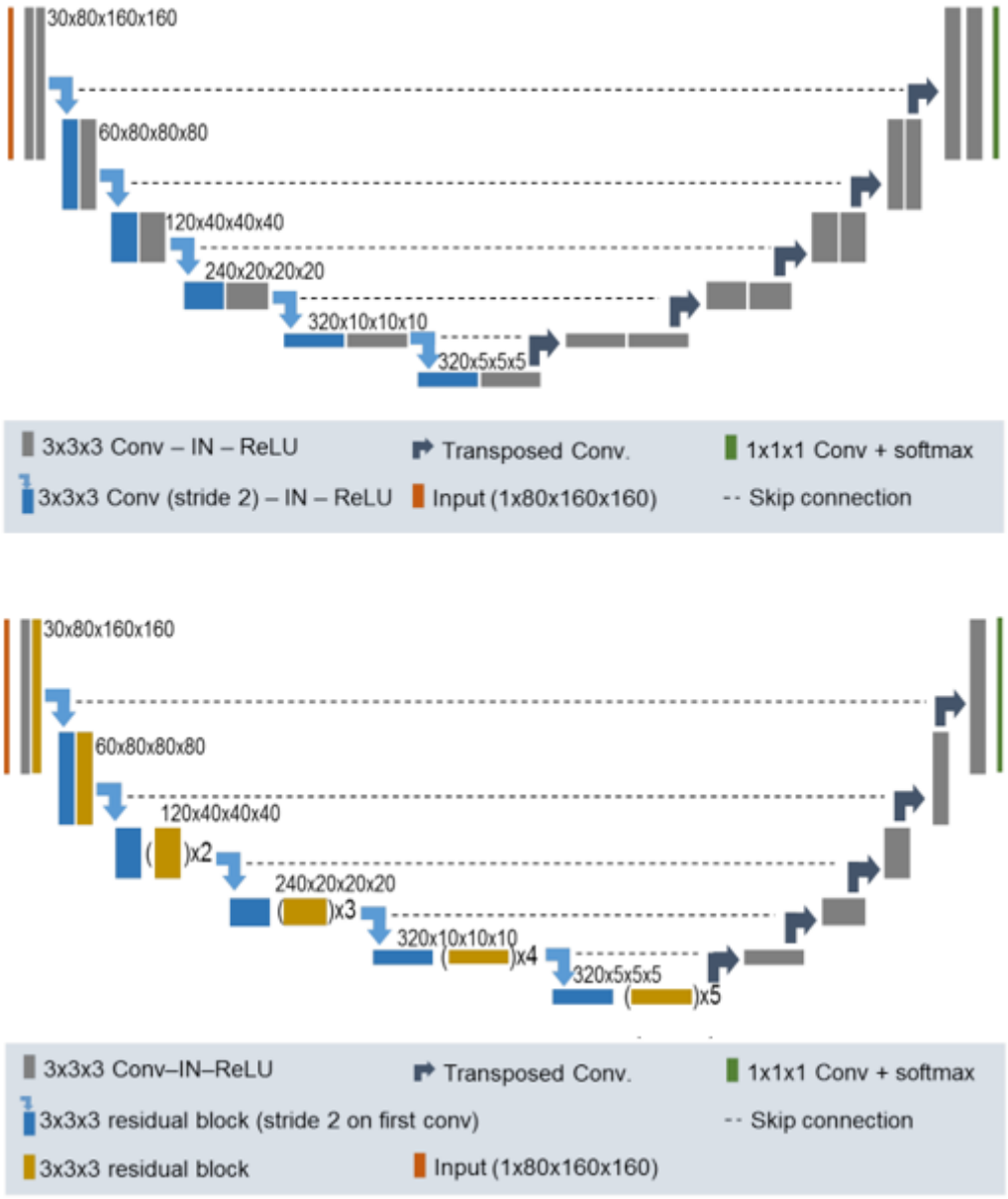

Figura 12. Esquemas de la arquitectura de Plain U-Net (arriba) y ResEnc U-Net (abajo), donde ()×X hace referencia al número de veces que se repite un bloque. Extraído de Gesthalter et al.[5].

4.2.2 Self-configuring U-Net: nn-Unet

El término *self-configuring* o autoconfiguración respecto a una U-Net hace referencia al proceso de automatización de todo el flujo de trabajo a la vez que se produce una adaptación del entrenamiento a las características del propio *dataset* para aumentar el rendimiento predictivo. Normalmente se cumple con varios requisitos o etapas principales para realizar una segmentación automática de tipo *self-configuring*:

- Primero se analiza automáticamente una descripción del *dataset*. Esta descripción general de las características del *dataset* creada por el usuario en formato .json, contiene parámetros como el tipo de imagen (como por ejemplo TC o RM) o el número de pacientes del estudio. A partir de este análisis o selección de características se puede determinar de manera automática aspectos como el tipo de normalización que se va a aplicar a las imágenes.
- Posteriormente, se analiza en profundidad el dataset en base a la caracterización anterior. Este procedimiento consiste en identificar características como el tamaño medio de la imagen tras el preprocesado o la intensidad media, conformando la huella del *dataset*. Esta "huella" es una descripción más exhaustiva de las imágenes de nuestro conjunto de datos, y pueden ser utilizados para aplicar transformaciones de preprocesado, calcular hiperparámetros en base a reglas y modificar las arquitecturas de las redes para que se adecúen a los datos y al hardware disponible.
- Finalmente, una vez realizado el entrenamiento, se evalúa automáticamente la configuración creada en base a las etapas anteriores. Si la evaluación no es satisfactoria se pueden cambiar ciertos parámetros y reentrenar todo el modelo.

En la actualidad, existen dos *frameworks* capaces de cumplir con todo este grado de automatización: Auto3DSeg (MONAI) [56] y nnU-Net [41], [42]. Finalmente se ha optado por nnU-Net debido a su mayor simplicidad, resultados y mayor apoyo por la comunidad científica.

El *framework* nnU-Net (no-new U-Net) se basa en la premisa de que no es necesario emplear arquitecturas modernas o complejas para obtener buenos resultados en la segmentación, sino que únicamente es necesario adaptar el proceso de segmentación a las propiedades del dataset y al hardware disponible. Para ello, establece una serie de parámetros que divide en parámetros fijos (*fixed*), parámetros basados en reglas (*rule-*

based) y parámetros empíricos (*empirical*). Los parámetros fijos se basan en una "plantilla" que se utiliza para todos los entrenamientos. En esta se fijan el *learning rate* o el optimizador entre otros. Los parámetros basados en reglas son aquellos que se calculan a partir de los datos aportados por la huella del dataset. Finalmente, los parámetros empíricos son aquellos que dependen del resultado del entrenamiento y las métricas de rendimiento para evaluar si son necesarios, como por ejemplo el post-procesado. En la Tabla 5 se recopilan todos los parámetros aplicables a un proceso de segmentación automática por medio de nnU-Net, agrupados por el tipo de parámetro que representan.

La combinación del cálculo y utilización de todos estos parámetros derivan en el siguiente flujo de trabajo:

1. Preprocesado y extracción de la huella del *dataset*
2. Cálculo de hiperparámetros y entrenamiento del modelo
3. Evaluación del modelo
4. Posible post-procesado

En la Figura 13 se observa cómo afecta a este flujo de trabajo al cálculo de los distintos parámetros.

Tabla 5. Parámetros presentes en una segmentación 2d mediante nnU-Net

Fixed parameters	Rule-based parameters	Empirical parameters
Learning rate	*Intesity normalization*	*Configuration of post-processing*
Loss function	*Image resampling strategy*	*Ensemble selection*
Architecture template	*Network topology*	
Optimizer	*Image target spacing*	
Data augmentation	*Batch size*	
Training procedure	*Patch size*	
Inference procedure	*Annotation resampling strategy*	

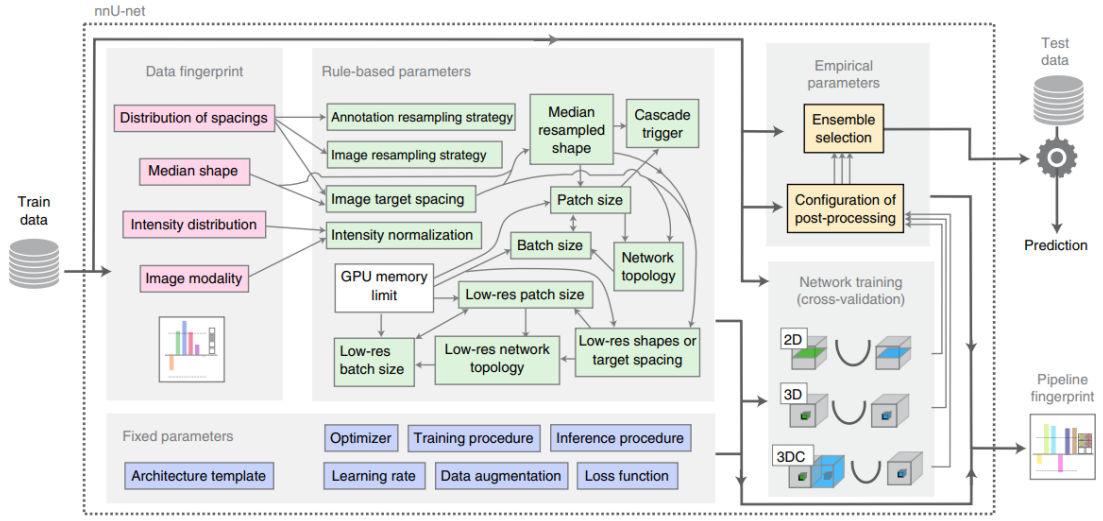

Figura 13. Flujo de trabajo de nnU-Net. Extraído de [41].

4.2.3 Flujo de trabajo nnU-Net

El flujo de trabajo es igual para las dos arquitecturas a comparar, pues únicamente difieren en la arquitectura propiamente dicha (combinación final de módulos y dimensiones). Este flujo de trabajo comienza caracterizando el *dataset* en un archivo .json en función de varios ítems.

- ***Channel_names*** nos permite darle un nombre al canal en el que se va a crear el modelo. Esto es habitual en el contexto de procesado de imagen. Para este *framework*, cualquier nombre distinto de "CT" realizará en el preprocesado una normalización de intensidad de tipo estandarización *zscore*.
- ***Labels*** especifica tanto el número de etiquetas como su nombre y el valor asociado a cada una.
- ***NumTraining*** establece el número de pacientes o conjuntos de imágenes que vamos a utilizar para el entrenamiento.
- ***File_ending*** y ***overwrite_image_reader_writer*** hacen referencia al formato en el que se encuentran almacenadas las imágenes y el lector de imagen que las va a cargar en memoria, respectivamente. En la Tabla 6 se observa el aspecto que tiene este archivo para nuestro estudio particular.

Tabla 6. Caracterización del dataset para su introducción el flujo de trabajo de Nn-Unet

Ítem	Valor
channel_names	0: zscore
labels	*background*: 0
	LB: 1
	RB: 2
	Ctraq: 3
numTraining	19
file_ending	.nii.gz
overwrite_image_reader_writer	SimpleITKIO

4.2.3.1 Preprocesado y extracción de la huella del *dataset*

Una vez caracterizado el *dataset*, lo primero que hay que hacer es crear una "huella" o caracterización general del mismo para, posteriormente, aplicar un preprocesado derivado de esta huella. Nn-Unet permite crear una huella a partir de los siguientes parámetros, que podrán ser utilizados o no para el preprocesado:

- **Valores mínimo y máximo del conjunto de imágenes**. Estos parámetros resultan de extrema utilidad para conocer el rango de intensidad que comprende al total de las imágenes. Según el origen de nuestras imágenes, éste debería ser [0, 255]. Con estos datos se pueden aplicar normalizaciones personalizadas a cada *dataset*. Estos valores se calculan como el argumento máximo o mínimo de cada imagen, para posteriormente identificar el argumento mínimo o máximo de todo el conjunto de argumentos. Su cálculo es muy común para todo tipo de señales y procesados [44], definido mediante las siguientes ecuaciones:

$$j_{max} = argmax_{j\epsilon\{1,\dots,M\}}\left(arg\max_{(x,y)} I_j(x,y)\right) \quad 4.1$$

$$j_{min} = arg\ min_{j\epsilon\{1,\dots,M\}}\left(arg\min_{(x,y)} I_j(x,y)\right) \quad 4.2$$

- **Media y mediana del conjunto de imágenes.** Este par de datos estadísticos sirven para caracterizar el valor de intensidad media y mediana, respectivamente, de cada imagen y posteriormente del conjunto total. Con estos valores, obtenidos mediante las siguientes ecuaciones se pueden realizar normalizaciones de intensidad, como la estandarización *zscore*.

$$\bar{I}(x,y) = \frac{1}{M}\sum_{j=1}^{M} I_j(x,y) \tag{4.3}$$

$$\tilde{I}(x,y) = median\big(I_1(x,y), I_2(x,y), \dots, I_M(x,y)\big) \tag{4.4}$$

- **Percentiles 0.5 y 99.5.** Estos percentiles permiten identificar *outliers* de intensidad dentro de las imágenes. Son muy utilizados en el procesado de imágenes de *CT* para hacer recortes en el histograma de intensidad [44].

- **Desviación estándar.** Este parámetro estadístico describe la dispersión de los datos respecto a su media. Se calcula con la finalidad de realizar una estandarización *zscore.*

- **Tamaño de todas las imágenes y su respectivo *spacing* después del recorte**. El flujo de trabajo de nnU-Net hace automáticamente un recorte en las dimensiones de cada paciente en la fase anterior al preprocesado en caso de ser necesario. Este recorte elimina zonas sin información como aire o *background* (común en imágenes de TC tridimensionales). Esto se describe también en la huella del *dataset* junto al número de imágenes con las que colabora cada paciente y su tamaño. El *spacing* hace referencia al tamaño del vóxel de cada paciente. Para imágenes 2D, como es nuestro caso, este parámetro será siempre 1. Tampoco sufriremos un recorte en las dimensiones por la misma razón.

A partir de esta huella se realiza el preprocesado, que consiste en una normalización de intensidad y en el cálculo de los parámetros basados en reglas. Este proceso se repite por cada arquitectura de red que se quiera probar. En nuestro caso vamos a comparar 2 arquitecturas distintas, por lo que este proceso se realizará dos veces.

La normalización de intensidad aplicada al *dataset* bajo estudio en el presente TFG fue la estandarización *zscore*. Esta estandarización consiste en la transformación de la

intensidad de las imágenes para que tengan media 0 y desviación estándar 1. Se ha elegido esta normalización de intensidad debido a que se ha utilizado y contrastado para imagen endoscópica en la detección de anormalidades como úlceras o pólipos [57]. Esta técnica se describe como el valor de intensidad menos la media de cada imagen entre la desviación estándar también cada nivel de imagen, de acuerdo a la ecuación (4.5). Con ello, se consigue que cada imagen individual tenga el mismo rango de intensidad y el mismo formato *float* para facilitar el entrenamiento.

$$I_{\text{std}}(x, y) = \frac{I(x, y) - \mu}{\sigma} \tag{4.5}$$

Una vez aplicado el preprocesado se crean varios archivos *.json*, que describen los planes de entrenamiento para el *dataset*. Además de esto, nnU-Net plantea una *five-fold cross-validation,* es decir, una validación cruzada de k=5 iteraciones, que divide el conjunto de entrenamiento en dos conjuntos, uno de entrenamiento y otro de validación, con una proporción de 80%-20% diferente en cada iteración. Esta división, que se utiliza para medir el rendimiento predictivo de cada permutación de datos, se almacena en un *.json* denominado *splits_final*. En la Tabla 7 se muestra la división de pacientes empleada en el entrenamiento y validación de los modelos propuestos en el presente TFG.

Los parámetros basados en reglas más relevantes para nuestro problema de estudio, calculados durante el preprocesado y contenidos en los planes del entrenamiento son: *(i) batch size, (ii) patch size y (iii) network topology*. Los demás parámetros solo se activan cuando la red va a ser tridimensional (nuestra red va a tener convoluciones bidimensionales) o cuando el *spacing* es distinto de 1, que tampoco es nuestro caso, pues en los vídeos de broncoscopia cada píxel mide lo mismo de alto que de ancho (540 x 540).

Tabla 7. Partición de entrenamiento y validación para cada fold de entrenamiento

	Fold 1	***Fold 2***	***Fold 3***	***Fold 4***	***Fold 5***
Entrenamiento					
	Paciente01	Paciente02	Paciente01	Paciente01	Paciente01
	Paciente02	Paciente03	Paciente02	Paciente02	Paciente04
	Paciente03	Paciente04	Paciente03	Paciente03	Paciente05
	Paciente05	Paciente05	Paciente04	Paciente04	Paciente09
	Paciente07	Paciente07	Paciente05	Paciente07	Paciente10
	Paciente09	Paciente09	Paciente07	Paciente09	Paciente11
	Paciente10	Paciente10	Paciente11	Paciente10	Paciente13
	Paciente13	Paciente11	Paciente13	Paciente11	Paciente14
	Paciente14	Paciente13	Paciente14	Paciente14	Paciente15
	Paciente15	Paciente15	Paciente16	Paciente15	Paciente16
	Paciente16	Paciente17	Paciente17	Paciente16	Paciente17
	Paciente19	Paciente20	Paciente19	Paciente17	Paciente19
	Paciente21	Paciente21	Paciente20	Paciente19	Paciente20
	Paciente22	Paciente22	Paciente22	Paciente20	Paciente21
	Paciente23	Paciente23	Paciente23	Paciente21	Paciente22
					Paciente23
Validación					
	Paciente04	Paciente01	Paciente09	Paciente05	Paciente02
	Paciente11	Paciente14	Paciente10	Paciente13	Paciente03
	Paciente17	Paciente16	Paciente15	Paciente22	Paciente07
	Paciente20	Paciente19	Paciente21	Paciente23	

Es imprescindible encontrar el equilibrio entre estos parámetros, para lograr un entrenamiento óptimo y realista de acuerdo a los recursos computacionales con los que disponemos. En nuestro caso, la GPU disponible es una RTX 4090 con 24GB de VRAM cedida por el Grupo de Ingeniería Biomédica (GIB) de la Universidad de Valladolid. A continuación, se describen los parámetros basados en reglas empleados y sus procedimientos de cómputo:

(i) ***Patch size*** o tamaño de parche. Hace referencia al tamaño de las imágenes que se van a procesar como input durante el entrenamiento. Puede ser de igual, mayor o menor que el tamaño de la imagen original, pero no se trata de un redimensionado, sino de un recorte que tiene por objetivo limitar el tamaño de la imagen remarcando zonas de interés para optimizar el proceso de entrenamiento minimizando la carga computacional. En general, tamaños de parche más grande permiten una mayor contextualización y rendimiento, que

se debe compensar con un tamaño de *batch* más pequeño. Estudios como el de Hamwood *et al.* [58] indican que un aumento del tamaño de parche aumenta el rendimiento en tareas de clasificación. Otros estudios, como el de Quintana *et al.* abogan por adaptar el tamaño del parche a la tarea y características del *dataset* [59]. Nn-Unet realiza justamente esto último. El *patch size* va a ser inicializado como el tamaño mediano de imagen del *dataset*, que se irá reduciendo iterativamente hasta que la red se pueda entrenar con un *batch size* de al menos 2, dependiendo de la capacidad de la GPU. Con esto conseguimos un compromiso entre *patch size* y *batch size* soportable por la GPU [41].

(ii) ***Batch size.*** Este parámetro hace referencia al número de imágenes que se van a procesar simultáneamente durante una iteración o época (*epoch*) del entrenamiento. En trabajos como el de Sato *et al.* se concluye que tamaños de *batch* grandes mejoran el rendimiento en tareas de segmentación respecto a entrenamientos con tamaños de *batch o patch* más pequeños [60]. Durante el entrenamiento, un mayor *batch* size permite una estimación del gradiente más precisa. Nn-Unet calcula este parámetro dependiendo del *patch size*. Si éste ha disminuido desde la inicialización del preprocesado, el *batch size* se fija en 2. Si no ha disminuido, la memoria de la GPU restante se utiliza para incrementar el tamaño de *batch* hasta que no quede más memoria libre. Sin embargo, este valor tiene un límite máximo que no aplica para imágenes bidimensionales [41].

(iii) ***Network topology.*** La topología de red hace referencia a la propia arquitectura estructural de la red. Describe cómo se organizan las capas y el número de niveles de profundidad, tanto en el *downsampling* como en el *upsampling*, así como el número de filtros o *features* durante la etapa codificadora y decodificadora. Nn-Unet adapta la topología de red a partir de una plantilla basada en la U-Net original. Ésta se calcula ajustando el número de operaciones de *downsampling* en cada eje (operaciones para reducir la resolución de la imagen capturando información a diferentes escalas) hasta que el tamaño de la imagen se reduzca a un mínimo de 4 píxeles desde el tamaño de *patch*. Además, el *kernel* de convolución es de 3×3 y con un *stride* o paso de 2 unidades se elimina la capa de *pooling* para disminuir la resolución. Por último, en la topología final habrá un total de $5 \times k + 2$ capas

convolucionales, donde *k* es el número de *downsampling operations* o reducciones de resolución [41].

4.2.4 Entrenamiento

El entrenamiento de las distintas redes es el proceso mediante el cual la red resuelve una tarea ajustando los pesos y sesgos de las neuronas y actualizándolos para que las salidas o predicciones se aproximen lo máximo posible a la salida deseada o *ground truth*.

Esto se consigue mediante la optimización de una función de pérdidas (*loss function*) que compara las predicciones con las etiquetas reales de manera iterativa, modificando los pesos para ajustarse a los datos reales. Para llevar a cabo un buen entrenamiento, primero hay que definir los hiperparámetros. Estos son el conjunto de decisiones iniciales que se adoptan antes de empezar el entrenamiento. Por defecto, nnU-Net tiene hiperparámetros preestablecidos. Sin embargo, debido a la experimentación y al apoyo de la literatura científica, se han variado algunos de estos parámetros para maximizar el rendimiento predictivo de las redes para este problema particular bajo estudio. Excluyendo el *patch size* y el *batch size*, que son calculados mediante el procedimiento de *self-tunning* descrito anteriormente, los hiperparámetros a determinar en esta etapa son los siguientes: el tipo de optimizador, con parámetros como *learning rate* inicial (LR) o tasa de aprendizaje y el planificador del LR; técnicas de regularización, como *weight decay, instnorm o dropout*; número de épocas y de iteraciones por época; función de pérdida o *loss function*; integración y tipo de *data augmentation*; *probabilistic oversampling*; función de activación; y *deep supervision*. Todos estos parámetros permanecen constantes durante el entrenamiento, por lo que su optimización previa puede dar lugar a resultados significativamente mejores.

- **Optimizador.** El optimizador es el algoritmo encargado de ajustar los pesos de la red durante el entrenamiento. Tiene por objetivo minimizar una función, la de pérdidas, de modo que las predicciones o salidas se parezcan lo máximo posible a las etiquetas reales. El proceso es el siguiente: primero, la red hace una predicción a partir de la entrada. Posteriormente, se calcula la función de pérdida comparando la predicción con el valor de la etiqueta real. A continuación, se usa *backpropagation* para calcular los gradientes, es decir, la dirección en la que se

van a ajustar los pesos y, finalmente, el optimizador es el que actualiza los pesos utilizando los gradientes previamente calculados. NnU-net utiliza por defecto el algoritmo SGD (*Stocastic Gradient Descent*), con un momento de Nesterov de μ = 0.99. Este es el optimizador original, muy usado en la actualidad e introducido por primera vez en 1951 por *Robbins* y *Monro* [61] y ampliado por *Kiefer* y *Wolfowitz* [62]. Este algoritmo se caracteriza por calcular el gradiente estimado a partir de muestras individuales o por lotes pequeños de datos elegidos aleatoriamente en vez de calcular el gradiente de todo el *dataset*. La ecuación que describe este proceso es la siguiente:

$$v_t = \alpha v_{t-1} - \eta \nabla f(\theta_t) \quad 4.6$$

siendo:

- α el factor de inercia (o momento)
- v_{t-1} la velocidad del paso anterior
- η el *learning rate* o la tasa de aprendizaje
- $\nabla f(\theta_t)$ el gradiente de la función de pérdida f en la posición θ_t.

Posteriormente Yurii Nesterov modificó este método, calculando el gradiente desde una posición anticipada $\theta_t + \alpha v_{t-1}$, para mejorar la rapidez de convergencia y la estabilidad del aprendizaje en funciones con cambios bruscos de gradiente [63].

$$v_t = \alpha v_{t-1} - \eta \nabla f(\theta_t + \alpha v_{t-1}) \quad 4.7$$

Para este TFG se ha utilizado el optimizador *Adan* o *Adaptative Nesterov Momentum Algorithm* [64]. Este optimizador se trata de una adaptación del método original de Nesterov que evita calcular el gradiente en un punto adelantado reduciendo el coste computacional y realiza una estimación adaptativa de los momentos de primer (media de los gradientes) y segundo (media de los cuadrados de los gradientes) orden del gradiente similar al optimizador *Adam*. Con estas modificaciones se consigue un tiempo de convergencia que se aproximaría al límite teórico conocido, siendo el actual estado del arte en optimizadores. Puede igualar o incluso mejorar a algunos optimizadores con la mitad de épocas de entrenamiento. A este optimizador le hemos asociado un *learning rate* de 0.001, común en la literatura [64].

- ***Learning rate* y planificador.** El *learning rate* o tasa de aprendizaje se define como la magnitud del cambio que se aplica en el ajuste de los pesos. Se puede asimilar a la distancia de los pasos que se dan hasta llegar a la meta o solución durante una carrera. Pueden ir en distintas direcciones y tener distintas longitudes, pero para llegar a la meta (converger) nos tenemos que colocar justo en el punto de la solución, no nos podemos pasar. Para lograr esto debemos establecer un *learning rate* inicial y una estrategia de variación de este parámetro para no "pasarnos" de la convergencia ni "quedarnos cortos". Como se ha indicado con anterioridad, el valor de la tasa de aprendizaje inicial es de 0.001, un valor pequeño. La estrategia de variación, o planificador que vamos a utilizar es el *PolinomialLRscheduler*. Este planificador varía la tasa de aprendizaje de manera lineal en función del número de épocas restantes, dando lugar a una curva de variación de la tasa de aprendizaje con forma de línea recta. No depende del rendimiento a diferencia de *ReduceLronPlateau* y es útil cuando se entrena un número fijo de épocas con cambios suaves.
- **Función de pérdidas (*loss*).** La función de pérdidas se trata de la función a optimizar durante el algoritmo de entrenamiento. Se define como la función que relaciona la predicción con la etiqueta real y cuyo error hay que minimizar. Existen numerosas funciones de pérdidas. Para la tarea de segmentación es bastante común encontrar combinaciones de funciones de pérdidas para crear una función global más robusta y que de mejores resultados que utilizar una única función [65], aunque no siempre es la mejor opción y muchas veces depende del *dataset* o de la tarea particular a realizar [66].

Para la presente tarea de segmentación se utilizó la combinación de las funciones de pérdida *BinaryCrossEntropy* (BCE) y *Dice Loss* (DC), denominada *Combo Loss.* Mientras que BCE evalúa la probabilidad de pertenencia de cada píxel a cada clase, DC mide la superposición entre la predicción y las etiquetas reales. Combinar ambas funciones facilita el aprendizaje de los límites de las segmentaciones optimizando la superposición general [65]. Además, NnU-Net cuenta con *deep supervision* (supervisión profunda). Esta es una técnica de entrenamiento en la que la función de pérdida se aplica tanto en la salida final de la red como en capas intermedias del decodificador, mejorando el flujo del gradiente y evitando el estancamiento. Las salidas y sus respectivas pérdidas tienen pesos distintos, de modo que cada una afecta

de forma distinta a la pérdida final. En nnU-Net el peso se va reduciendo a la mitad por cada disminución de resolución. Finalmente, cabe destacar que, si bien el valor idóneo para una pérdida BCE es de 0, y para DC es de 1, al combinar ambas la "mejor" pérdida posible es de -1, en vez de 0.

- **Técnicas de regularización: *weight decay*, *instnorm* y *dropout*.** Las técnicas de regularización son estrategias orientadas a disminuir el sobreajuste (*overfitting*) en el modelo final, es decir, que la red memorice y no sea capaz de generalizar. Esto resulta un problema importante, ya que una red con sobreajuste puede alcanzar buenos resultados durante el entrenamiento, pero no será capaz de extrapolar ese buen rendimiento a datos con los que no haya sido entrenada. No se habrán aprendido los patrones generales de los datos, sino que se habrá memorizado características particulares del conjunto concreto de entrenamiento.

Para nuestra tarea de segmentación hemos empleado distintas técnicas para evitar este problema:

- *Weight decay* o decaimiento de pesos. Es una técnica de regularización aplicable por medio del optimizador durante el proceso de entrenamiento. Consiste en reducir gradualmente el valor de los pesos, como si se desvanecieran a lo largo del tiempo. Esto se aplica sobre todo si los pesos contribuyen en menor medida al resultado final. El valor que se le ha dado a *weight decay* es de 0.00003, un valor muy pequeño para no alterar demasiado los valores de los pesos, y utilizado comúnmente en la literatura [67].
- *InstNorm*. Es el bloque que realiza la técnica de normalización por instancias. Tiene por objetivo estabilizar el entrenamiento y mejorar la generalización y es independiente del tamaño del *batch*. La normalización por instancias o normalización de contraste, de uso específico para tareas con imágenes, sustituye el *batch normalization*, una forma de normalización más comúnmente utilizada [68]. Se diferencia de esta ya que se aplica de forma individual para cada imagen y canal en vez de por lotes. Se calcula la media y desviación de cada imagen y canal y se normaliza cada imagen mediante estandarización *zscore.*

- *Dropout*. Es una técnica de regularización que consiste en "apagar" o desactivar sistemáticamente algunas conexiones entre neuronas de la capa oculta de manera aleatoria. Con esto se consigue no memorizar y mejorar la generalización. En la documentación de *Pytorch* se describe como marcar como 0 algunos de los elementos del *input* de la red con una probabilidad *p*. Si bien NnU-Net y sus creadores no contemplan esta estrategia como adecuada, ya que empíricamente no han obtenido buenos resultados, para este trabajo se ha marcado un *dropout* = 0.4, es decir, un 40% del input será 0.
- *Early stopping*. Es una técnica de regularización muy comúnmente utilizada para terminar el entrenamiento de manera temprana si la red no está obteniendo buenos resultados en términos de la optimización de la función de *loss*. En el presente trabajo no se ha aplicado en base a las recomendaciones de los creadores de nnU-net, ya que empíricamente no encontraron mejoras significativas en su uso. En cualquier caso, durante el procedimiento de aprendizaje se van guardando los pesos actualizados cuando disminuye el *loss*, por lo que, aunque no se termine el entrenamiento, se almacenan siempre los mejores resultados.

- ***Probabilistic oversampling*.** El sobremuestreo probabilístico es una técnica de muestro adaptativo que pretende mejorar la precisión de la segmentación para estructuras infrarrepresentadas en el *dataset*. Esta técnica consiste en aumentar la probabilidad de selección de muestras o regiones durante el entrenamiento que contengan clases minoritarias de manera controlada en base a una probabilidad escogida. Para la presente tarea de segmentación, la carina traqueal se encuentra ligeramente infrarrepresentada. En términos relativos, supone un 30% de la representación total de las clases, mientras que el bronquio izquierdo y el bronquio derecho ocupan alrededor de un 36.5% y 33.5% respectivamente. Por ello, se ha marcado una probabilidad de sobremuestreo de 0.66 (66%).
- **Función de activación**. La función de activación es una parte de la arquitectura cuyo objetivo es determinar si una neurona debe ser o no activada, favoreciendo la introducción de no linealidad en el modelo. Gracias a ello, la red puede aprender relaciones más complejas y mejorar la aproximación de funciones arbitrarias. La función de activación empleada en este TFG ha sido Leaky ReLU, con pendiente igual a 0.01 en base a las recomendaciones de los creadores de nnU-Net basados en *Andrew*

et al. [69]. Esta función de activación es especialmente útil, ya que permite el flujo del gradiente aunque la entrada sea negativa, algo que la ReLU convencional no permitía. Además, resulta de utilidad para evitar saturaciones y mejora la propagación del gradiente. En la Figura 14 se puede observar la forma de esta función de activación. La ecuación que describe esta curva es la siguiente:

$$f(x) = \begin{cases} x, & \text{si } x \geq 0 \\ \alpha x, & \text{si } x < 0 \end{cases} \quad 4.8$$

siendo α la pendiente del primer tramo de la recta, marcada en 0.01 para nuestro estudio.

- **Número de épocas e iteraciones por época.** El número de épocas hace referencia al número de iteraciones que se van a realizar durante el proceso de entrenamiento. Por defecto, nnU-net entrena con 1000 épocas. Sin embargo, debido a los cambios anteriores y al gasto computacional, el tiempo destinado a realizar un entrenamiento de 1000 épocas era totalmente inasumible. Con una duración media de 7 minutos por época, realizar un entrenamiento completo supondría alrededor de 117 horas de computación intensiva. Con la finalidad de reducir esta carga, se ha marcado un número de épocas arbitrario de 150. Sin embargo, para algunos *folds* de la arquitectura Plain U-Net debido a la alta varianza de pérdida y métricas de validación, se ha parado manualmente el entrenamiento antes de completar las 150 épocas. Se espera que este

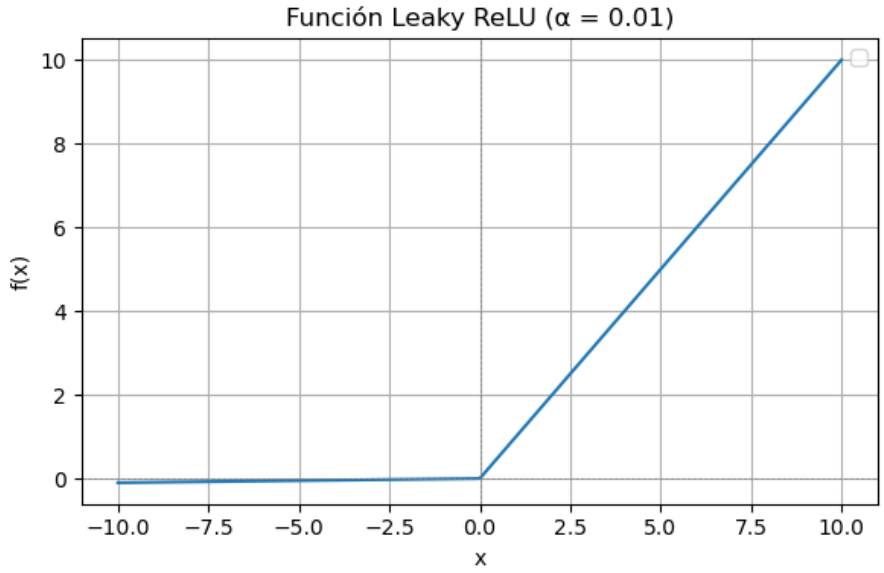

Figura 14. Gráfica representativa de la función de activación Leaky ReLU.

"*early stopping* manual" no afecte al resultado final, puesto que como se ha comentado con anterioridad, los pesos de la red son almacenados cada época que mejora la función de *loss*. Comprobaciones *a posteriori* de la forma de la gráfica de entrenamiento confirmaron la idoneidad de esta estrategia.

El número de iteraciones por época o mini-*batches*, tanto de entrenamiento como de validación, hacen referencia al número de volúmenes/imágenes que se procesan por cada época. Los parámetros por defecto son 250 para el entrenamiento y 50 para la validación. Esto quiere decir que por cada época se procesan 250 × *batch size* imágenes o volúmenes durante el entrenamiento, calculándose las métricas de validación (que indican si el entrenamiento puede ser generalizable o no) con 50 × *batch size* imágenes/volúmenes. Modificar estos valores resulta de especial utilidad cuando se dispone de pocos datos de entrenamiento, como es nuestro caso. Con 19 pacientes para el entrenamiento, 250 iteraciones por época no son suficientes imágenes a procesar para lograr unos resultados robustos. Es por esto por lo que se ha utilizado un total de 1500 iteraciones por época para el entrenamiento, aumentando consecuentemente hasta 100 el número de iteraciones por época para la validación. Para una *batch size* hipotético de 2, en vez de procesar 500 imágenes por época (250×2) se procesarían 3000 imágenes (1500×2). Si complementamos este aumento de imágenes con *data augmentation* somos capaces de obtener unos resultados notablemente mejores a si no modificásemos este parámetro en absoluto.

Data augmentation. La técnica de *data augmentation* se caracteriza por la generación de imágenes sintéticas o artificiales mediante el procesado de imágenes reales del *dataset* que se utilizarán como datos de entrada durante el entrenamiento. La generación de imágenes y sus respectivas etiquetas en nnU-Net se crean "*on the fly*", es decir, se generan y almacenan durante cada época y posteriormente se eliminan para no ocupar memoria. La técnica *data augmentation* está conformada por distintas transformaciones, como rotaciones, difuminados o recortes, entre otras. El propio *framework* tiene disponibles varias opciones y niveles pre-establecidos de "agresividad" en las transformaciones. Esto hace referencia a la probabilidad de que sucedan estas transformaciones en el contexto bajo estudio y a lo mucho o poco que modifican la imagen original. Para la tarea de segmentación propuesta en este TFG, se ha adoptado una postura agresiva respecto a esta etapa, debido a la falta de datos suficientes para poder generalizar. En la Tabla 8 se muestran todas las transformaciones realizadas asociadas a su probabilidad de ocurrencia,

así como a otros parámetros de interés relativos a cada transformación. Además, en la Figura 15 se pueden observar estas transformaciones.

Tabla 8. Tabla representativa de las transformaciones de data augmentation más relevantes llevadas a cabo durante el entrenamiento

Transformación	**Probabilidad**	**Parámetros asociados**
SpatialTransform	1	
• Elasticdeform	0.5	alpha: (85,125), sigma: (8,12)
• Rotation	0.8	
• Scale	0.5	scale: (0.7,1.43);
Rot90Transform	0.4	
MedianFilterTrasform	0.2	kernel size: (2,8)
GaussianBlurTransform	0.25	sigma: (0.8, 1.5)
GaussianNoiseTransform	0.25	
BrightnessTransform	0.4	range: (0, 0.5)
ContrastAugmentationTransform (preserve_range=True)	0.4	range: (0.5, 2)
ContrastAugmentationTransform (preserve_range=False)	0.4	range: (0.5, 2)
SimulateLowResolutionTransform	0.1	zoom: (0.8, 1); order: down=0, up=3
GammaTransform	0.1	
MirrorTransform	1 (aleatorio)	
BlackRectangleTransform	0.05	n: (1, 5) size: [p//10, p//3]
BrightnessGradientAdditiveTransform	0.2	range: (-0.5, 1.5)
LocalGammaTransform	0.2	range: (-0.5, 1.5)
SharpeningTransform	0.2	strength: (0.1, 1)

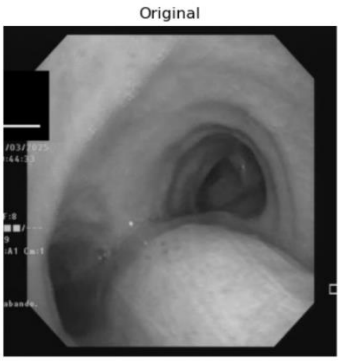

Figura 15. Imagen original sobre la que se aplica data augmentation.

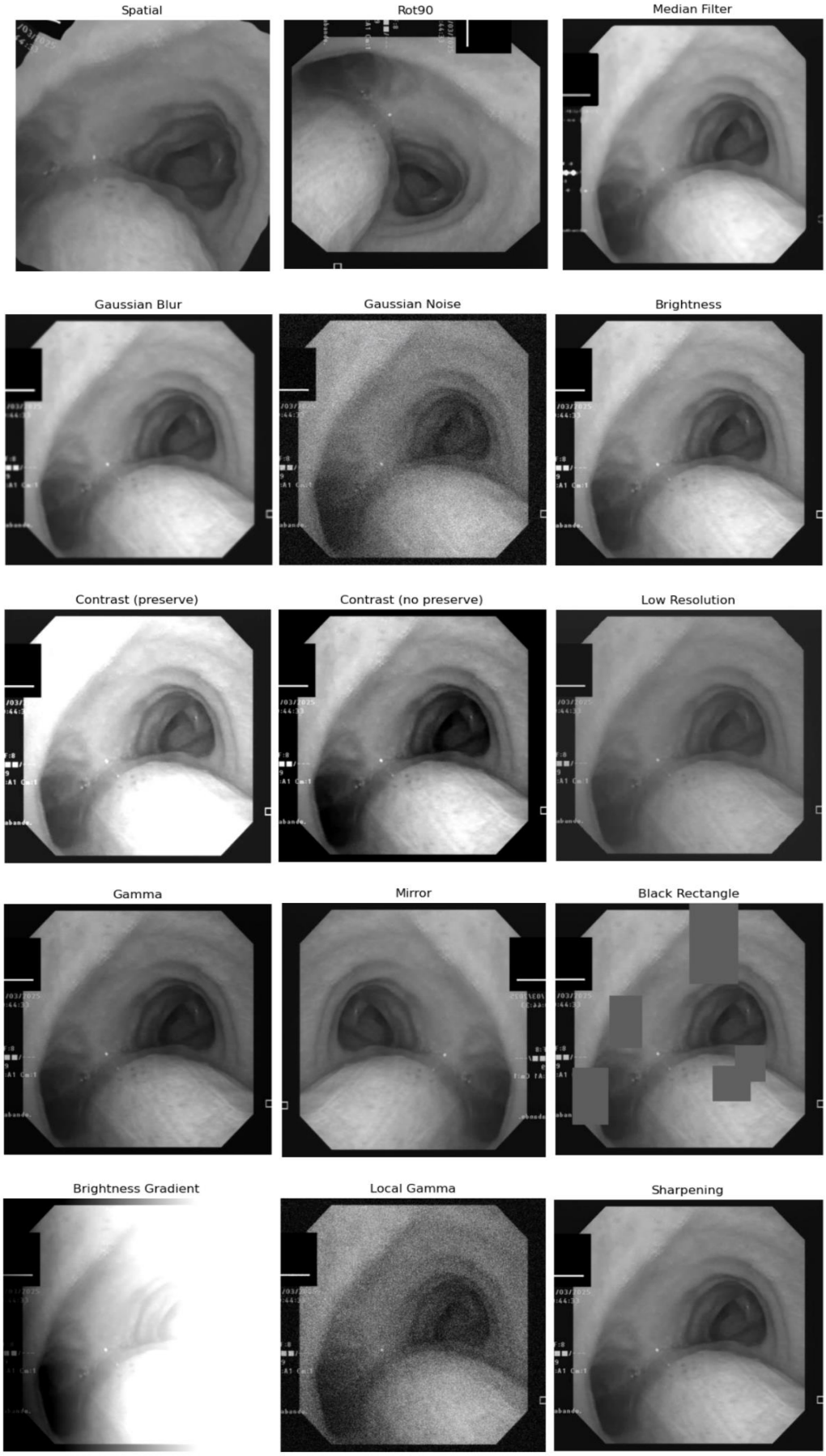

Figura 16. Transformaciones de data augmentation sobre la imagen original.

Finalmente, a modo de resumen y para facilitar la lectura, la Tabla 9 muestra una recopilación de todos los hiperparámetros aplicados durante el entrenamiento en el presente TFG.

Tabla 9. Recopilación de los hiperparámetros de entrenamiento.

Hiperparámetro	
Optimizador	Adan
Learning Rate	1e-03
LR *scheduler*	PolyLRScheduler
Técnicas de regularización	SI
Weight decay	3e-05
InstNorm	SI
Dropout	0.4
Número de épocas	150
Número de iteraciones	
Entrenamiento	1500
Validación	100
Loss function	BCE + DC (Combo loss)
Deep supervision	SI
Probabilistic oversampling	0.66
Función de activación	Leaky ReLU
Data augmentation	Agresiva

4.3 Postprocesado y localización

Tras analizar el rendimiento predictivo del modelo, surge la duda de si estos resultados se pueden mejorar. El postprocesado en segmentación de imagen médica es el conjunto de técnicas, algoritmos y transformaciones aplicados a la salida de la red para mejorar la robustez de las soluciones, simplificar las predicciones y lograr maximizar los parámetros de rendimiento descritos con anterioridad. De hecho, el postprocesado es uno de los parámetros empíricos de la nnU-Net. El *framework* primero calcula las métricas de rendimiento. Si queremos aplicar el postprocesado, el nnU-Net aplica su estrategia de postprocesado basado en componentes conectados y evalúa el resultado de las métricas de rendimiento aplicadas sobre las predicciones postprocesadas [41]. Comparando las métricas de las imágenes sin procesar contra las procesadas, ofrece *feedback* sobre si el postprocesado es necesario o empeora los resultados [41]. En este trabajo se van a evaluar dos estrategias de postprocesado distintas: la aplicada automáticamente por nnU-Net, basada en componentes conectados, y un postprocesado basado en el suavizado gaussiano de las segmentaciones; comparando cuál de ellas es más efectiva para el problema de estudio.

- Postprocesado de componentes conectados. Es un algoritmo utilizado comúnmente en la segmentación de imagen médica, sobre todo para la segmentación de órganos como hígado, riñón y tumores de riñón [74], [75]. Consiste en eliminar todos los elementos de un mismo valor o etiqueta menos el más grande o de mayor área/volumen. Con esto se consigue eliminar falsos positivos distribuidos por zonas de la imagen lejanas al *ground truth*. En la Figura 17 podemos visualizar el efecto de este postprocesado para la mejora del rendimiento. Eliminando todos los componentes con la etiqueta de bronquio derecho menos el de mayor área mejoramos las métricas de rendimiento durante la evaluación ya que los demás componentes son falsos positivos.
- Postprocesado basado en el suavizado gaussiano. El postprocesado por suavizado gaussiano consiste en aplicar un kernel de convolución gaussiano de modo que se produzca un difuminado. Al aplicar este kernel, la matriz de predicción cambia su tipo de bit de entero (uint8) a *float*, aumentando el contorno de las segmentaciones con valores artificiales. Si establecemos un umbral del valor mínimo, vamos a introducir estos valores en la nueva segmentación de manera controlada y reconvirtiendo la matriz a uint8, obtenemos como resultado unos bordes de

segmentación más regulares y lisos, deseables en nuestra tarea de segmentación. Un ejemplo de este postprocesado se describe en la Figura 18. Para no alterar demasiado los bordes de la segmentación, se ha establecido un *sigma* de 2 y un umbral mínimo de 0.4 de manera empírica comprobando de manera visual un suavidad aceptable. Técnicas similares de postprocesado de imágenes a partir de filtros gaussianos se han utilizado para mejorar la segmentación de imagen médica, obteniendo buenos resultados en términos del coeficiente DSC [76].

Finalmente, una vez optimizada la segmentación de las estructuras del árbol bronquial, es posible plantear la localización de la posición del broncoscopio dentro de la propia anatomía. Esto se consigue en base a las segmentaciones realizadas, aplicando una asociación del contexto de las etiquetas con las estructuras segmentadas. En la Tabla 11 se muestra esta asociación, donde en función de las estructuras segmentadas se interpreta la localización en un lugar u otro del árbol bronquial.

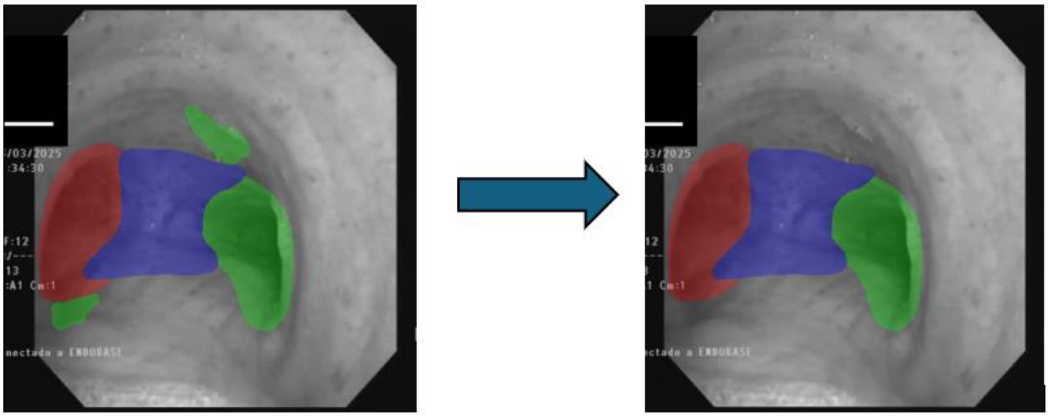

Figura 17. Postprocesado basado en componentes conectados.

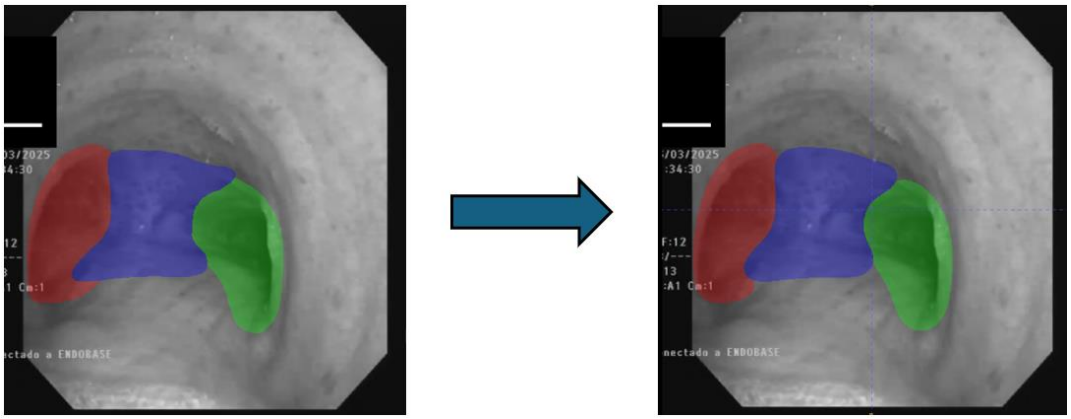

Figura 18. Postprocesado basado en filtrado gaussiano

A continuación, se describe la estrategia de localización implementada en el presente TFG. En primer lugar, durante la segmentación de una imagen se enumeran todas las estructuras segmentadas. Por ejemplo, para la primera bifurcación, la red tiene que haber marcado algún píxel como *background*, LB, RB y CT. Al estar presentes todas estas estructuras en una misma imagen, podemos inferir que estamos en un punto de la anatomía donde somos capaces de ver estas tres estructuras, es decir, la primera bifurcación. Lo mismo sucede si en la imagen se segmentan las estructuras *background*, LB y CT. Al no haber segmentado RB, se entiende que esta estructura no se encuentra presente en la imagen y, por tanto, el broncoscopio se encuentra orientado hacia el bronquio izquierdo o incluso dentro de este.

Los denominados artefactos son predicciones erróneas o imposibles en el contexto anatómico, que solo se pueden dar en las predicciones. Un ejemplo puede ser observar en la misma imagen el bronquio derecho y el izquierdo sin observar la carina traqueal. Debido a que son estructuras contiguas, es imposible que suceda si la segmentación del *ground truth* ha sido correctamente realizada. Para las etiquetas reales no debería haber artefactos, mientras que para las predicciones sí que puede haberlos.

Con este método se consigue una localización generalista y poco específica por estructuras. La evaluación del rendimiento de la localización bronquial se realizará por medio de una matriz de confusión multiclase, ya descrita como parte de las métricas de rendimiento para la segmentación. Esta matriz cuenta con varias clases enfrentadas entre sí dependiendo de su condición de valor real o predicción. Una correspondencia total entre la etiqueta anatómica de la predicción y la etiqueta anatómica del *ground truth* contabilizará como un verdadero positivo. La falta de correspondencia se considerará como una asociación falsa que se comparará con otras clases, para evaluar en qué clases fallan más estas predicciones. Finalmente, se extraerán métricas derivadas de esta matriz de confusión, como el valor predictivo positivo (*precision*) y la sensibilidad (*recall*).

4.4 Evaluación y métricas de rendimiento

Una vez definida la estrategia de entrenamiento, es preciso definir cómo se va a evaluar el grado de aprendizaje y qué métricas de rendimiento se van a utilizar para ello. Esta evaluación se divide en dos partes: validación durante el propio entrenamiento y evaluación después del entrenamiento.

*Tabla 10. Asociación de estructuras segmentadas con su etiqueta anatómica. * hace referencia a que cualquier conjunto de valores de predicción que no contenga 0 será marcado como artefacto.*

Elemento	Estructuras segmentadas	Valores de predicción	Etiqueta Anatómica
1º	LB + RB + CT	[0,1,2,3]	Primera Bifurcación
2º	LB + CT o LB	[0,1,3] o [0,1]	Grupo Bronquio Izq.
3º	RB + CT o RB	[0,2,3] o [0,2]	Grupo Bronquio Dcho.
4º	Artefacto	[0], [0,3], [0,1,2] *	-

La evaluación durante el entrenamiento comprende todas las métricas de rendimiento calculadas durante el propio entrenamiento sobre el conjunto de validación para observar el progreso de la red. Las métricas presentes durante el entrenamiento son el *pseudo dice* y el EMA pseudo dice. Estos parámetros se calculan en cada iteración (*epoch*) del entrenamiento.

- ***Pseudo dice***. El *pseudo dice* hace referencia al *Dice Score* convencional (explicado más adelante). Sin embargo, debido a que nnU-Net trabaja con parches en vez de con imágenes enteras y a que hacer una validación real de imágenes completas ralentizaría significativamente el entrenamiento; el *pseudo dice* es el cálculo del *Dice Score* de parches aleatorios tomados del conjunto de validación pegados entre sí, pretendiendo que fueran una imagen o volumen completo. Este parámetro únicamente se utiliza para ver cómo evoluciona el entrenamiento y no pretende dar una estimación o métrica precisa de cómo terminará siendo el entrenamiento. Si existen varias clases para segmentar se calcula el pseudo dice para cada clase.
- **EMA pseudo dice**. *Exponential Moving Average pseudo dice*, o EMA *pseudo dice* es una métrica que, al igual que el *pseudo dice*, no pretende dar una estimación precisa, sino medir visualmente cómo evoluciona el entrenamiento. Para minimizar el impacto de la variación del *pseudo dice*, EMA *pseudo dice* realiza una media móvil exponencial de los *pseudo dice* calculados anteriormente mediante la siguiente ecuación:

$$updated = old * alpha + (1 - alpha) * new \qquad 4.9$$

donde:

- *old* y *new* hacen referencia a los *pseudo dices* anteriores o actuales (si hay varias etiquetas se estima la media)
- *alpha* es el factor exponencial

Con ello, se consigue una curva más suave, que ayuda a representar mejor la verdadera tendencia del entrenamiento.

La evaluación tras el entrenamiento es la que de verdad nos indica el grado de precisión de la red que hemos creado y cómo de bien cumple con la tarea para la que ha sido creada. El *framework* de nnU-Net facilita la evaluación de los modelos con una plantilla de evaluación prefabricada. Sin embargo, esta se puede modificar fácilmente para albergar más métricas de rendimiento de las que ya se encuentran incorporadas oficialmente. Para esta trabajo, se han evaluado las siguientes métricas de rendimiento: métricas de la matriz de confusión (FN, FP, TP, TN), métricas basadas en F-*score* (IoU y DSC), HD95, *precision*, *recall*, N_pred y N_ref.

- Métricas de la matriz de confusión. Debido a que la tarea propuesta se trata de una segmentación semántica de imágenes, es decir, evaluar si cada píxel se ha segmentado correctamente, es necesaria una métrica que evalúe si la predicción respecto al píxel es verdadera o falsa. Una matriz de confusión es una matriz que describe la relación entre predicciones y etiquetas reales o *ground truth*, de forma que le asigna una "valor" a una relación entre predicción y etiqueta real. Este valor puede ser verdadero positivo (TP) si la predicción del píxel segmentado es de la misma clase que la etiqueta real, verdadero negativo (TN) si predicción y etiqueta son *background*, falso positivo (FP) si la predicción no se corresponde con la etiqueta siendo la etiqueta *background* y falso negativo (FN) donde la predicción es *background* pero la etiqueta contenía una segmentación real. Cuando existen varias clases, como es nuestro caso, se evalúa clase por clase y las clases que no están siendo evaluadas cuentan como *background* [70], [71]. En la Tabla 10 se muestra cómo se organiza una matriz de confusión.

Tabla 11. Matriz de confusión

		PREDICCIONES	
		VERDADERO	NEGATIVO
ETIQUETAS REALES	VERDADERO	Verdadero positivo (TP)	Falso negativo (FN)
	NEGATIVO	Falso positivo (FP)	Verdadero negativo (TN)

A partir de estos parámetros se pueden calcular varias métricas de rendimiento.

- Métricas F-*score*: IoU y *Dice Similarity Coefficient*. Las métricas basadas en el parámetro F-*score* son de las más utilizadas en el entorno de la visión por computadora para medir el rendimiento en tareas de segmentación. Se calculan a partir de la sensibilidad y la precisión (valor predictivo positivo) de las predicciones, que representan el grado de solapamiento entre la predicción y el *ground truth* o segmentación real, penalizando los falsos positivos. Existen dos métricas principales basadas en el parámetro F-*score*. Estas son el índice de Jaccard o *Intersection-over-Union* (IoU) y el F1-score o *Dice similarity coefficient* (DSC), que además es conocido como la media armónica entre precisión (i.e., valor predictivo positivo) y sensibilidad, siendo el más aceptado y usado por la comunidad científica [70]. Ambos índices se calculan de la siguiente manera:

$$IoU = \frac{TP}{TP + FP + FN} \tag{4.10}$$

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{4.11}$$

- Precision o valor predictivo positivo (PPV). Esta métrica describe la relación entre los píxeles clasificados correctamente como píxeles verdaderos y la suma de los píxeles distintos a *background* ya sean verdaderos o falsos. Esta métrica se suele utilizar como parte de otras métricas. Por sí sola hace referencia a la precisión de

la segmentación, es decir, cuantos píxeles marcados como positivos son correctos. Se calcula mediante la siguiente ecuación:

$$Precision = \frac{TP}{TP + FP} \quad 4.12$$

- *Accuracy* o exactitud. La métrica *accuracy* describe la tasa de aciertos del modelo. Se define como la proporción de predicciones correctas sobre el total de predicciones y se utiliza para caracterizar cómo de bien predice la red en término de aciertos absolutos. Esta métrica se calcula mediante la siguiente ecuación:

$$Accuracy = \frac{TP\ +\ TN}{TP\ +\ TN\ +\ FP + FN} \quad 4.13$$

- Recall o sensibilidad. La sensibilidad, *recall* o tasa de verdaderos positivos, mide la proporción de píxeles positivos reales correctamente detectados por la red. Se complementa con el valor predictivo positivo, de manera que el PPV determina cuantos de los píxeles que marca la red como positivos son ciertos, mientras que el *recall* estima de todos los positivos cuántos ha logrado detectar el modelo. Se calcula mediante la siguiente ecuación:

$$Recall = \frac{TP}{TP + FN} \quad 4.14$$

- N_pred y N_ref. N_pred se define como la suma total de píxeles que el modelo ha predicho como positivos. Se calcula como la suma entre los verdaderos positivos y los falsos positivos. En contraposición a esta métrica, n_ref hace referencia al número de píxeles que realmente pertenecen a la clase evaluada. Entre las dos dan una idea aproximada del rendimiento del modelo. Se calculan de la siguiente manera.

$$npred = TP + FP \quad 4.15$$

$$nref = TP + FN \quad 4.16$$

- HD95 - Hausdorff *distance* (95%). La distancia de Hausdorff es una métrica ampliamente utilizada basada en la distancia espacial como medida de similitud [72]. Se recomienda su empleo sobre todo cuando la *accuracy* de la segmentación es de especial importancia en el contexto del estudio. Representa la distancia mínima entre dos conjuntos de puntos (predicción y *ground truth*) y se focaliza en la evaluación de la delineación de las fronteras [70]. Debido a su sensibilidad por los *outliers*, HD95 ignora el 5% de los peores datos calculados, dando lugar a una métrica más estable. Así como las métricas anteriores se presuponen mejores cuanto mayor es su valor, la distancia de Hausdorff clasifica una segmentación como mejor que otra cuanto menor es su valor. Se calcula mediante la siguiente ecuación:

$$\mathrm{d}(\mathrm{A},\mathrm{B}) = \frac{1}{N}\sum_{\mathrm{a}\in\mathrm{A}}^{\mathrm{n}} \min_{b\in B} \left|\left|a-b\right|\right| \qquad 4.17$$

Todas las métricas anteriores ayudan a caracterizar la calidad de las segmentaciones en mayor o menor medida. Sin embargo, DSC destaca por ser la más utilizada y aceptada en el ámbito de la segmentación de imagen médica. Con valores de DSC de al menos 0.7 o superior, tradicionalmente ya se considera como aceptable una segmentación [73]. Además, para segmentaciones multiclase, cuantas más etiquetas tenga el *dataset*, por lo general menor será el DSC asociado a cada etiqueta. El objetivo será entonces alcanzar un DSC lo más cercano a 0.7 o superior, con la finalidad de obtener segmentaciones fiables y aceptables por la comunidad.

5. Resultados

En este capítulo se describen los resultados experimentales obtenidos tras la aplicación de la metodología descrita anteriormente.

Partiendo de 35 vídeos de broncoscopia recopilados de manera prospectiva, se ha llegado a segmentar hasta la primera bifurcación de todos ellos. Inicialmente, se planteó una división 70-30 entre entrenamiento y test de la población de estudio, común en el ámbito del aprendizaje computacional. Sin embargo, la incorporación de nuevos pacientes debido al carácter prospectivo de la fase de reclutamiento una vez diseñadas y entrenadas las redes de segmentación propuestas (Plain U-Net y ResEnc U-Net) , obligó a la incorporación de estos nuevos pacientes al grupo de test. Si bien esto podría limitar el aprendizaje de la red, puede servir para realizar una validación más robusta de los resultados. Finalmente, la proporción fue de 55-45, con 19 pacientes de entrenamiento y 16 de test.

Los resultados se enfocan en la comparación de las distintas arquitecturas de red y sus respectivos rendimientos predictivos. Además, también se han comparado las estrategias de postprocesado, para determinar cuál de las dos resulta de mayor utilidad para la tarea de segmentación de imágenes de broncoscopia. Finalmente, se ha evaluado el rendimiento predictivo de la localización en el árbol bronquial mediante imágenes segmentadas por parte de las dos arquitecturas.

5.1 Resultados del entrenamiento

Para llevar a cabo el entrenamiento de cada arquitectura con el *framework* nnU-Net, existen elementos o parámetros compartidos por ambas estrategias (parámetros fijos), es decir, la mayor parte de los hiperparámetros o la división de pacientes entre entrenamiento y validación; así como otros parámetros basados en reglas que se calculan a partir de las características particulares del *dataset*. Debido a la gestión de los recursos de memoria y a que cada arquitectura ocupa un espacio distinto en el disco, estos parámetros, además de variar en función del *dataset*, variarán también por cada tipo de arquitectura propuesta.

Tabla 12. Tabla comparativa de los parámetros basados en reglas de las arquitecturas propuestas.

Arquitectura	Plain U-Net	ResEnc U-Net
Patch size	640×640	512×512
Batch size	8	12
Número de capas	8	8
Número de filtros por capa	[32, 64, 128, 256, 512, 512, 512]	[32, 64, 128, 256, 512, 512, 512]
Tamaño medio de la red (en GB)	0.905	2.735

En la Tabla 12 podemos apreciar una comparativa del cálculo de estos parámetros para cada arquitectura. Es posible observar cómo Plain U-Net tiene un tamaño de parche mayor incluso que el tamaño mediano de las imágenes (640×640 frente a los 540×540 originales). Sin embargo, para ResEnc U-Net el tamaño de *patch* es menor que el original (512×512 frente a los 540×540 originales). Esto se debe principalmente al compromiso *patch size – batch size* y al distinto espacio que ocupa cada arquitectura. El *framework* optimiza la memoria de la GPU de modo que se logre el máximo contexto posible con el máximo tamaño de *batch*. Para Plain U-Net de menos peso, se prioriza el contexto, incluso aplicando *padding* (relleno de 0 en los bordes de la imagen) para un *batch size* de 8. Para ResEnc U-Net, al ser una arquitectura más pesada y ocupar más memoria en la GPU, se va alterado este equilibrio y es necesario disminuir el contexto dividiendo las imágenes, pero aumentado el *batch size*.

5.1.1 Plain U-Net

Para la arquitectura Plain U-Net, se evaluó la adecuación del proceso de entrenamiento de manera gráfica, inspeccionando visualmente la evolución de las Figuras XX (a)-(f). En primer lugar, se entrenaron los 5 *folds*, cada uno con particiones de datos entrenamiento – validación distintas. Con ello se consigue evaluar cómo se comporta el modelo con distintos grupos de datos. Finalmente, se muestra la gráfica correspondiente al entrenamiento con todos los pacientes de entrenamiento y sin grupo de validación (esta se podría tomar de muestras aleatorias de cualquier paciente), con el objetivo de obtener un modelo único. Este es el modelo final que se ha evaluado en el conjunto de test independiente.

Cabe destacar que el entrenamiento ha durado aproximadamente 250 segundos por época (4 minutos y 16 segundos), por lo que un entrenamiento completo de 150 épocas comprendió unas 10 horas y 20 minutos. Como se ha comentado en la sección de metodología, si bien no se ha implementado *early stopping*, para Plain U-Net varía el número de épocas. En el *fold* 0 se ha alargado el entrenamiento para comprobar si las métricas de validación eran capaces de mejorar. Para los siguientes *folds* se ha limitado a 150 épocas o menos, dependiendo de la variación de los resultados. El *fold all* también se ha limitado a 150 épocas. En este caso, las métricas de validación no describen una evolución del entrenamiento "real", puesto que las imágenes de validación se toman de los mismos pacientes con los que se ha realizado el entrenamiento.

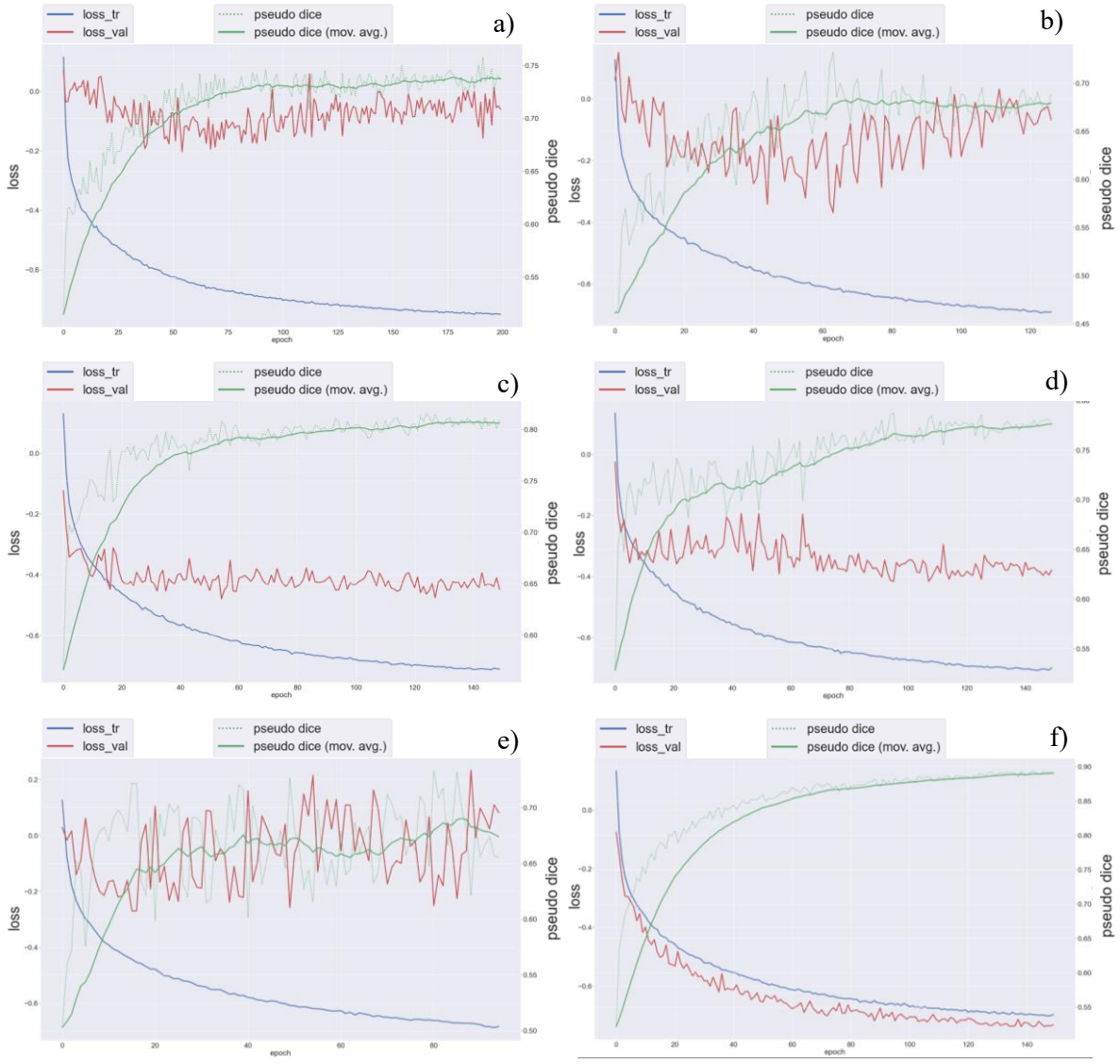

Figura 19. Comparativa de gráficas de entrenamiento para la arquitectura Plain U-Net y los distintos folds de entrenamiento: folds (a) 0, (b) 1, (c) 2, (d) 3, (e) 4 y (f) all.

Al evaluar las imágenes, los entrenamientos de los *folds* 0, 1 y 4 muestran signos de *overfitting*. Además, sus métricas de *pseudo dice* son bastante pobres. Para todos los *folds* hay bastante variabilidad entre épocas, sin embargo, esta variabilidad es muy notable para el *fold* 4. El *fold* 2 es el que a priori obtiene mejores resultados, tanto de la función de pérdida como de *pseudo dice*. En general, los resultados son bastante sólidos, y en el *fold all* no existe tanta variabilidad. Esto podría ser consecuencia del limitado número de sujetos de los que se obtienen las imágenes de entrenamiento.

5.1.2 ResEnc U-Net

El entrenamiento para la arquitectura ResEnc U-Net se muestra gráficamente a través de las Figuras XX (a)-(f), de la misma manera que para la arquitectura Plain U-Net. En primer lugar, se muestran todos los *folds* con las mismas divisiones de pacientes entrenamiento – validación, para terminar con una última figura que muestra el entrenamiento *fold all*, donde todos los pacientes se utilizan como entrenamiento.

Para esta arquitectura, el tiempo medio de entrenamiento por época ha sido de alrededor de 420 segundos. Todos los entrenamientos han estado limitados a 150 épocas, por lo que el tiempo de entrenamiento medio ha sido de 17 horas y 30 minutos.

Al evaluar las imágenes, se aprecia más variabilidad que la encontrada en la arquitectura Plain U-Net. Además, en la mayoría de los *folds*, el valor del *loss* de validación tiende a aumentar en vez de disminuir. Esto puede indicar que la red no está aprendiendo correctamente y que existe un *overfitting* bastante pronunciado. De nuevo, esto podría tener su origen en el limitado número de pacientes originales, a pesar de la aplicación de técnicas agresivas de *data augmentation*.

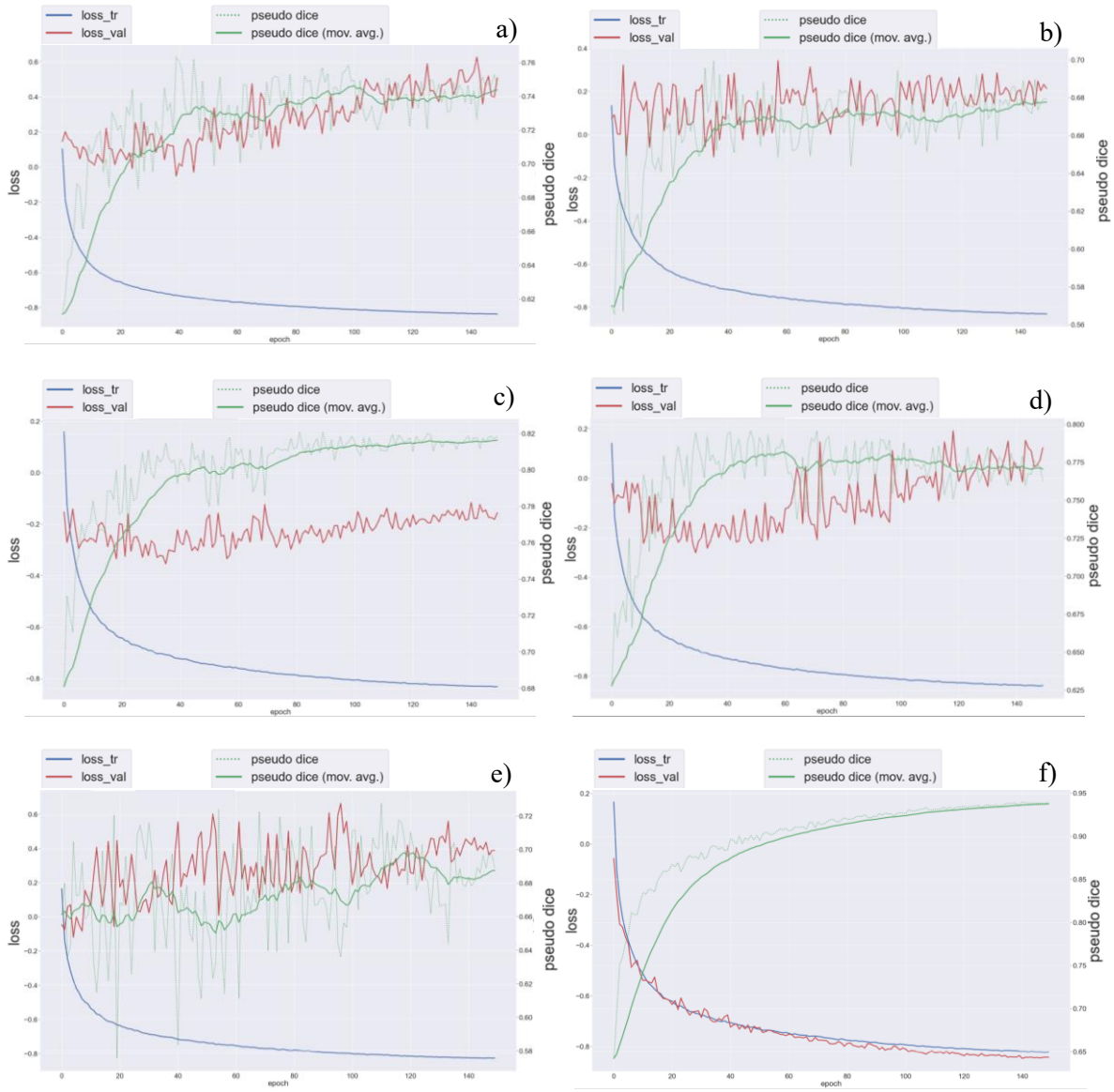

Figura 20. Comparativa de gráficas de entrenamiento para la arquitectura ResEnc U-Net y los distintos folds de entrenamiento: folds (a) 0, (b) 1, (c) 2, (d) 3, (e) 4 y (f) all.

5.2 Validación en el conjunto independiente de test

Para comprobar la calidad de las segmentaciones realizadas por las redes se utilizaron las métricas de rendimiento descritas en la metodología. Se han extraído las métricas de cada red original, con los resultados sin procesar, con los resultados postprocesados mediante la estrategia de componentes conectados y con los resultados postprocesados mediante la estrategia de filtrado gaussiano con sigma 2 y umbral de 0.4.

5.2.1 Plain U-Net

Las métricas de rendimiento de la red sin postprocesar se muestran en las Tablas 13 y 14. A la vista de los resultados obtenidos, se puede comprobar como todas las clases superan el 70% de DSC. La clase peor segmentada es la carina traqueal. Sin embargo, esta clase también se considera como aceptable. Las distancias HD95 son relativamente bajas, los parámetros *precision* y *recall* son bastante elevados. El índice IoU, sin embargo, es todavía mejorable.

Para la estrategia de postprocesado basado en componentes conectados, las métricas de rendimiento se muestran en las Tablas 15 y 16. La tabla es exactamente igual a la de los parámetros de los resultados sin procesar. Esto se debe a que la red consigue segmentar las tres clases como un único elemento conectado, por lo que no tiene que eliminar falsos positivos aislados, algo positivo para una segmentación multiclase.

Las métricas de rendimiento de la red tras el postprocesado basado en suavizado gaussiano se muestran en las Tablas 17 y 18. Los resultados para este tipo de postprocesado son, en general, peores, debido a la modificación de los límites de las segmentaciones. Sin embargo, se puede apreciar una mejora sustancial para el parámetro *N pred*, y su consecuente impacto en la métrica *recall*, que mejora en alguna clase respecto al anterior postprocesado. Respecto a la matriz de confusión, aumenta ligeramente el número de verdaderos positivos, aunque aumenta también el número de falsos positivos.

En la Figura 21 podemos comprobar gráficamente la similitud de las segmentaciones según los distintos enfoques. La predicción y ambos postprocesados dan lugar a imágenes casi idénticas. Visualmente el parecido de las segmentaciones al *ground truth* resulta satisfactorio. Para los distintos postprocesados a simple vista no se aprecian diferencias significativas.

Tabla 13. Métricas de rendimiento de la matriz de confusión en el conjunto de test para la arquitectura Plain U-Net sin postprocesado.

Etiqueta	Valor	TN*	TP*	FN*	FP*
LB	1	6.88	0.504	0.0817	0.125
RB	2	6.79	0.573	0.0838	0.138
CT	3	6.78	0.490	0.159	0.154
Foreground	[1,2,3]	6.82	0.523	0.108	0.139

* Valores de número de píxeles $x10^7$.

Tabla 14. Métricas de rendimiento derivadas de la matriz de confusión en el conjunto de test para la arquitectura Plain U-Net sin postprocesado.

Etiqueta	Valor	DSC	HD95	IoU	Precision	Recall	N pred	N ref
LB	1	0.835	7.631	0.719	0.801	0.881	6.30e+06	5.86e+06
RB	2	0.841	6.155	0.729	0.818	0.874	7.11e+06	6.57e+06
CT	3	0.766	9.908	0.623	0.779	0.763	6.45e+06	6.50e+06
Foreground	[1,2,3]	0.814	7.898	0.69	0.799	0.839	6.62e+06	6.31e+06

Tabla 15. Métricas de rendimiento de la matriz de confusión en el conjunto de test para la arquitectura Plain U-Net con procesado de componentes conectados.

Etiqueta	Valor	TN*	TP*	FN*	FP*
LB	1	6.88	0.504	0.0817	0.125
RB	2	6.79	0.573	0.0838	0.138
CT	3	6.78	0.490	0.159	0.154
Foreground	[1,2,3]	6.82	0.523	0.108	0.139

* Valores de número de píxeles $x10^7$.

Tabla 16. Métricas de rendimiento derivadas de la matriz de confusión en el conjunto de test para la arquitectura Plain U-Net. con procesado de componentes conectados.

Etiqueta	Valor	DSC	HD95	IoU	Precision	Recall	N pred	N ref
LB	1	0.835	7.631	0.719	0.801	0.881	6.30e+06	5.86e+06
RB	2	0.841	6.155	0.729	0.818	0.874	7.11e+06	6.57e+06
CT	3	0.766	9.908	0.623	0.779	0.763	6.45e+06	6.50e+06
Foreground	[1,2,3]	0.814	7.898	0.69	0.799	0.839	6.62e+06	6.31e+06

Tabla 17. Métricas de rendimiento de la matriz de confusión en el conjunto de test para la arquitectura Plain U-Net con procesado de suavizado gaussiano.

Etiqueta	Valor	TN*	TP*	FN*	FP*
LB	1	6.87	0.501	0.0846	0.133
RB	2	6.78	0.574	0.0825	0.155
CT	3	6.76	0.498	0.152	0.175
Foreground	[1,2,3]	6.80	0.525	0.106	0.155

* Valores de número de píxeles $x10^7$.

Tabla 18. Métricas de rendimiento derivadas de la matriz de confusión en el conjunto de test para la arquitectura Plain U-Net. con procesado de suavizado gaussiano.

Etiqueta	Valor	DSC	HD95	IoU	Precision	Recall	N pred	N ref
LB	1	0.825	8.033	0.704	0.788	0.873	6.35e+06	5.86e+06
RB	2	0.831	6.706	0.714	0.799	0.874	7.29e+06	6.57e+06
CT	3	0.76	11.952	0.615	0.756	0.774	6.73e+06	6.50e+06
Foreground	[1,2,3]	0.805	8.897	0.677	0.781	0.84	6.79e+06	6.31e+06

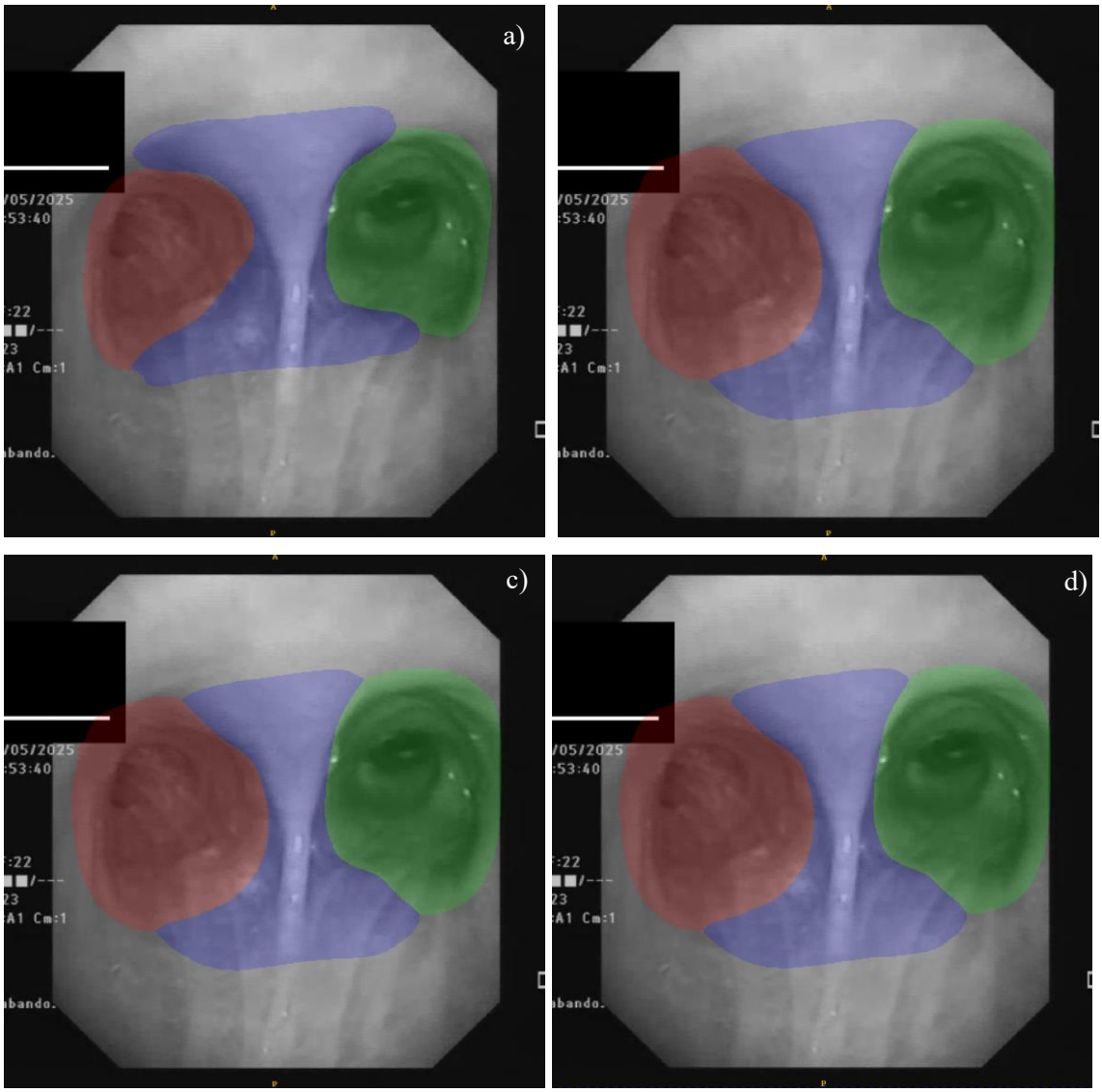

Figura 21. Ejemplo de imágenes segmentadas a partir de la arquitectura Plain U-Net donde: a) Ground Truth; b) Predicción base, c) Predicción con postprocesado de componentes conectados; d) Predicción con postprocesado de suavizado gaussiano.

5.2.2 ResEnc U-Net

Las métricas de rendimiento de la red sin procesar se muestran en las Tablas 19 y 20. Los resultados de esta arquitectura son bastante similares a los de la arquitectura anterior. Para algunas clases, la distancia HD95 es algo menor. La métrica *precision* es algo más elevada para esta arquitectura, aunque para la métrica *recall* se comporta inversamente.

Las métricas de rendimiento de la red tras el postprocesado basado en componentes conectados se muestran en las Tablas 21 y 22. Para esta arquitectura, al igual que para la arquitectura anterior, el procesado basado en componentes conectados resulta de poca

utilidad, puesto que las métricas permanecen intactas. Esto nos indica, de nuevo, que no existen falsos positivos aislados y que la red es capaz de segmentar por estructuras compactas.

Las Tablas 23 y 24 resumen las métricas de rendimiento de la red tras el postprocesado basado en suavizado gaussiano. En comparación con la arquitectura sin postprocesado, los resultados en general han empeorado notablemente. Sucede lo mismo que para la otra arquitectura. Al modificar los bordes de las segmentaciones, se incrementa el número de verdaderos positivos, pero también el de falsos positivos.

En la Figura 22 podemos comprobar gráficamente la similitud de las segmentaciones según los distintos enfoques del mismo *frame* y paciente que se muestran para la red anterior. La predicción y ambos postprocesados en este caso varían de los resultados obtenidos con la arquitectura anterior, aunque visualmente el parecido de las segmentaciones al *ground truth* también resulta satisfactorio. Para esta arquitectura la carina traqueal es más estrecha y pierde parte de su característica forma de "I". Entre los distintos tipos de postprocesado no se aprecian diferencias significativas visualmente.

Tabla 19. Métricas de rendimiento de la matriz de confusión en el conjunto de test para la arquitectura ResEnc U-Net sin postprocesado.

Etiqueta	Valor	TN*	TP*	FN*	FP*
LB	1	6.91	0.482	0.104	0.0958
RB	2	6.82	0.556	0.101	0.113
CT	3	6.79	0.458	0.192	0.145
foreground	[1,2,3]	6.84	0.499	0.132	0.118

* Valores de número de píxeles $x10^7$.

Tabla 20. Métricas de rendimiento derivadas de la matriz de confusión en el conjunto de test para la arquitectura ResEnc U-Net sin postprocesado.

Etiqueta	Valor	DSC	HD95	IoU	Precision	Recall	N pred	N ref
LB	1	0.827	5.914	0.708	0.838	0.825	5.78e+06	5.86e+06
RB	2	0.841	6.434	0.728	0.836	0.853	6.69e+06	6.57e+06
CT	3	0.737	11.59	0.586	0.78	0.716	6.03e+06	6.50e+06
foreground	[1,2,3]	0.802	7.979	0.674	0.818	0.798	6.17e+06	6.31e+06

Tabla 21. Métricas de rendimiento de la matriz de confusión en el conjunto de test para la arquitectura ResEnc U-Net con procesado de componentes conectados.

Etiqueta	Valor	TN*	TP*	FN*	FP*
LB	1	6.91	0.482	0.104	0.0958
RB	2	6.82	0.556	0.101	0.0113
CT	3	6.79	0.458	0.192	0.145
Foreground	[1,2,3]	6.84	0.499	0.132	0.118

* Valores de número de píxeles $x10^7$.

Tabla 22. Métricas de rendimiento derivadas de la matriz de confusión en el conjunto de test para la arquitectura ResEnc U-Net. con procesado de componentes conectados.

Etiqueta	Valor	DSC	HD95	IoU	Precision	Recall	N pred	N ref
LB	1	0.827	5.914	0.708	0.838	0.825	5.78e+06	5.86e+06
RB	2	0.841	6.434	0.728	0.836	0.853	6.69e+06	6.57e+06
CT	3	0.737	11.59	0.586	0.78	0.716	6.03e+06	6.50e+06
Foreground	[1,2,3]	0.802	7.979	0.674	0.818	0.798	6.17e+06	6.31e+06

Tabla 23. Métricas de rendimiento de la matriz de confusión en el conjunto de test para la arquitectura ResEnc U-Net con procesado de suavizado gaussiano.

Etiqueta	Valor	TN*	TP*	FN*	FP*
LB	1	6.90	0.480	0.106	0.104
RB	2	6.80	0.557	0.0999	0.128
CT	3	6.78	0.466	0.184	0.163
Foreground	[1,2,3]	6.83	0.501	0.130	0.131

* Valores de número de píxeles $x10^7$.

Tabla 24. Métricas de rendimiento derivadas de la matriz de confusión en el conjunto de test para la arquitectura ResEnc U-Net. con procesado de suavizado gaussiano.

Etiqueta	Valor	DSC	HD95	IoU	Precision	Recall	N pred	N ref
LB	1	0.816	7.17	0.691	0.824	0.816	5.84e+06	5.86e+06
RB	2	0.832	7.066	0.715	0.818	0.855	6.84e+06	6.57e+06
CT	3	0.734	14.539	0.582	0.76	0.728	6.29e+06	6.50e+06
Foreground	[1,2,3]	0.794	9.592	0.663	0.8	0.8	6.33e+06	6.31e+06

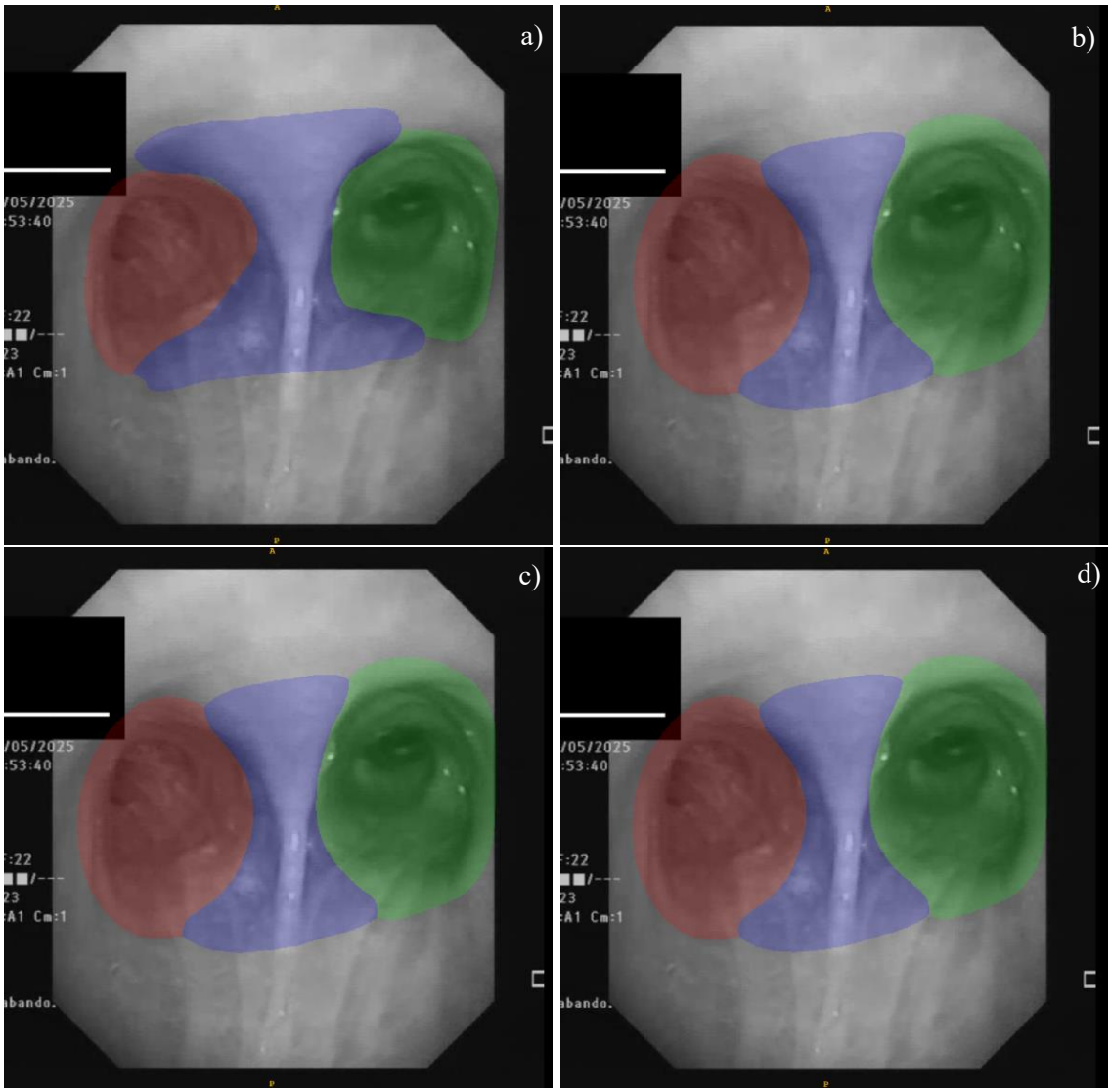

Figura 22. Ejemplo de imágenes segmentadas a partir de la arquitectura ResEnc U-Net donde: a) Ground Truth; b) Predicción base, c) Predicción con postprocesado de componentes conectados; d) Predicción con postprocesado de suavizado gaussiano.

5.3 Localización

La localización bronquial sería la aplicación final de este trabajo en la práctica clínica. Debido a que las redes sin postprocesar son las que han obtenido mejores resultados en cuanto a sus métricas de segmentación, estas son las que han sido utilizadas para evaluar la tarea de localización. Los resultados se muestran en forma de matriz de confusión multiclase, una para cada arquitectura. Debido a que el *ground truth* por definición no debería tener artefactos, estos, en caso de existir, se han eliminado de la matriz de confusión.

La Figura 23 muestra la matriz de confusión para la arquitectura Plain U-Net, en la que se comparan tres clases (primera bifurcación, grupo bronquio izquierdo y grupo bronquio derecho) y la parte de artefactos.

En la matriz de confusión se puede apreciar cómo el Grupo RB es el que peor identificado se encuentra. Además, destaca el gran porcentaje de errores confundiendo al grupo RB con la primera bifurcación. El número de artefactos es muy limitado y la red identifica como estas anomalías las imágenes que deberían pertenecer al grupo RB. Tanto el Grupo LB como la primera bifurcación han obtenido buenos resultados. En la Tabla 25 se observan las métricas de rendimiento para esta matriz de confusión.

Es posible observar cómo los valores de la matriz de confusión se corresponden con métricas de rendimiento aceptables. Se aprecia cómo el Grupo RB es el peor identificado en todas las métricas, teniendo un desempeño peor para la sensibilidad o el *recall*.

Para la arquitectura ResEnc U-Net, la matriz de confusión se muestra en la Figura 24.

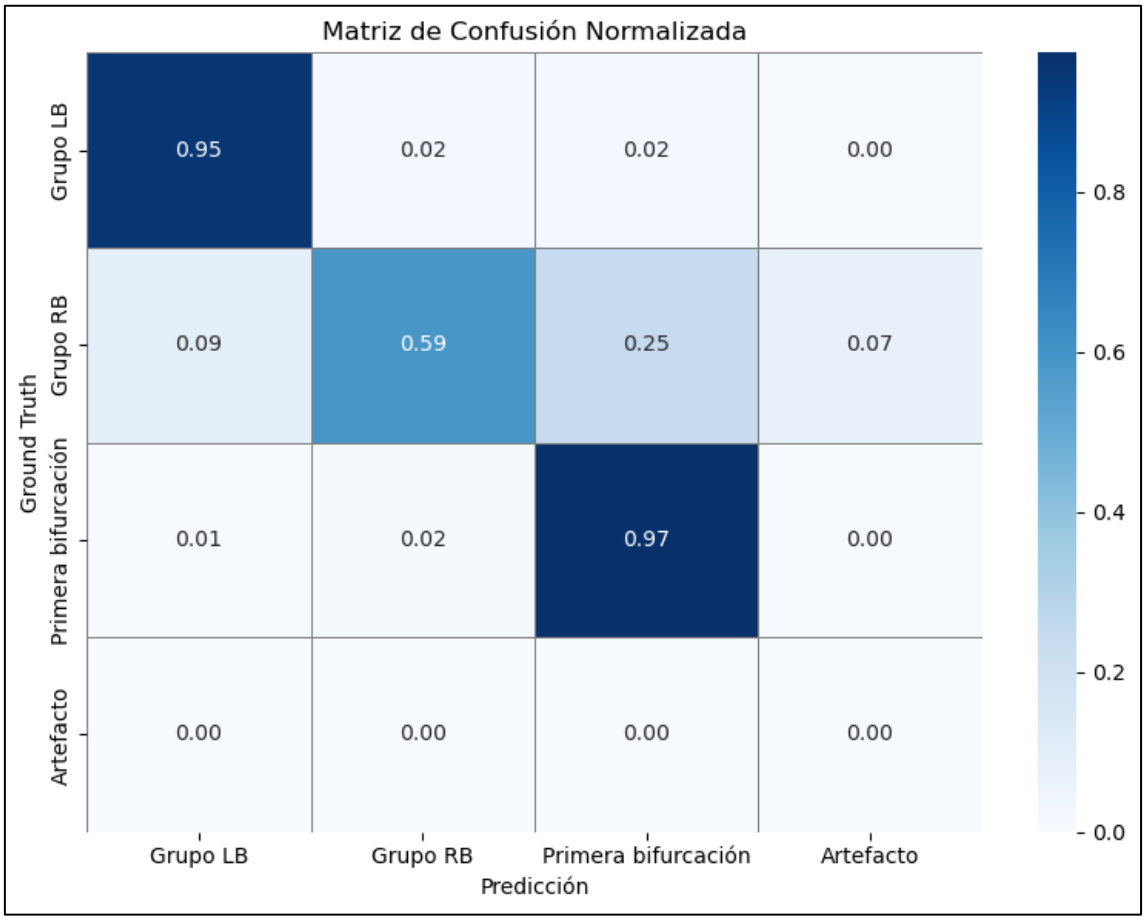

Figura 23. Matriz de confusión para la tarea de clasificación para la arquitectura Plain U-Net.

Tabla 25. Tabla de métricas de rendimiento asociadas a la matriz de confusión para la arquitectura Plain U-Net

Zona anatómica	PPV	*Recall*	*F1 - score*	*Accuracy*
Grupo LB	0.92	0.95	0.94	0.95
Grupo RB	0.67	0.59	0.63	0.59
Primera bifurcación	0.98	0.97	0.98	0.97

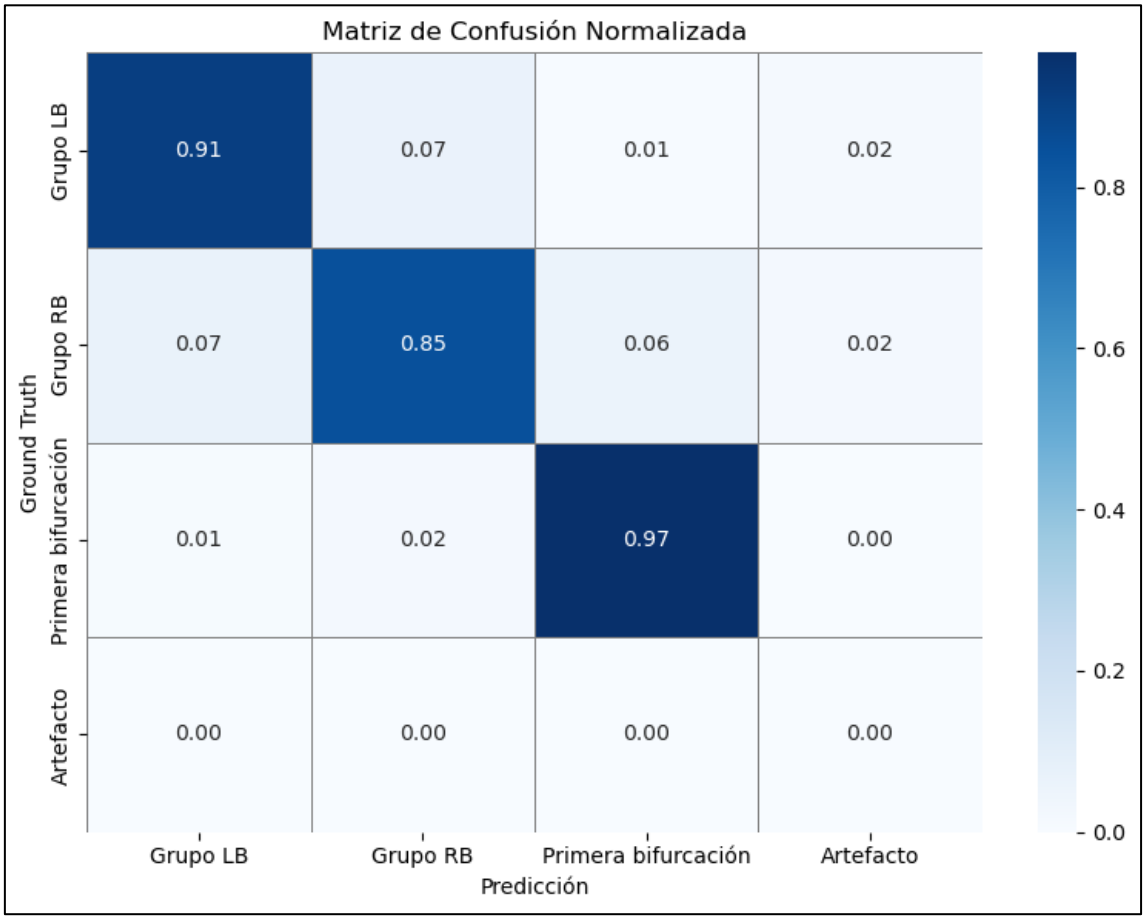

Figura 24. Matriz de confusión para la tarea de clasificación para la arquitectura ResEnc U-Net

Se puede apreciar una mejoría notable respecto a la anterior arquitectura para el Grupo RB. Para esta, no solo aumenta el número de imágenes del Grupo RB identificadas como tal, sino que, además, el número de artefactos es menor. Las métricas de rendimiento para esta matriz de confusión se muestran en la Tabla 26.

Se aprecia también una mejoría notable para el Grupo RB, mejorando en todas sus métricas y alcanzando niveles bastante más aceptables.

Finalmente, analizando el *ground truth* en total, se han reportado 9 imágenes artefactadas. El aspecto de estas imágenes se puede observar en las Figuras 25 (a) y (b). Son combinaciones de etiquetas imposibles, ya que, como se aprecia en la imagen, no se

ha segmentado la carina traqueal, cuando ésta claramente se encuentra entre el bronquio derecho (verde) y el bronquio izquierdo (rojo), o se ha guardado una imagen sin segmentación asociada debido a un error al realizar la misma.

Tabla 26. Tabla de métricas de rendimiento asociadas a la matriz de confusión para la arquitectura ResEnc U-Net.

Zona anatómica	VPP	Recall	F1 - score	Accuracy
Grupo LB	0.9	0.91	0.9	0.91
Grupo RB	0.65	0.85	0.74	0.85
Primera bifurcación	0.99	0.97	0.98	0.97

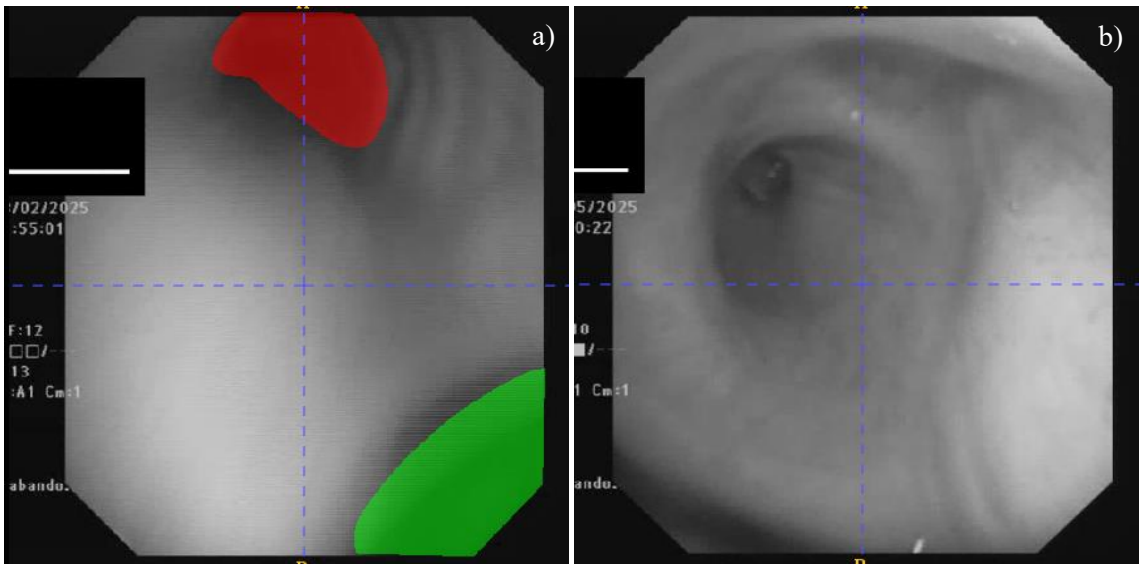

Figura 25. Imágenes artefactadas del ground truth: a) falta la carina traqueal; b) no se ha guardado la segmentación asociada a la imagen.

6. Discusión

Durante este trabajo se ha creado una herramienta de localización bronquial a partir de imágenes segmentadas del árbol bronquial, además de la respectiva herramienta de segmentación. Durante este TFG se ha desarrollado y validado prospectivamente un modelo de segmentación automática de las estructuras principales de la vía aérea (carina traqueal, bronquio derecho y bronquio izquierdo) basado en arquitecturas de deep learning aplicadas a imágenes de broncoscopia, que además se ha evaluado como herramienta de localización simplificada de la sonda del broncoscopio dentro del árbol bronquial. Para ello, se ha dividido la metodología de trabajo en varias etapas fundamentales: (i) creación del dataset, (ii) modelo de segmentación automática y (iii) localización endobronquial.

6.1 Creación del dataset

Durante esta etapa, se han transformado los vídeos de broncoscopia recopilados en la unidad de técnicas del Servicio de Neumología del Hospital Universitario Río Hortega de Valladolid en matrices de datos homogéneas y estandarizadas aptas para el análisis automático. Se han procesado de acuerdo a las estrategias de los estudios más recientes de análisis de imágenes endoscópicas y se han organizado para adaptarlas a un formato común para imágenes de TC, que es el estándar más empleado

De los 35 pacientes recopilados, se han segmentado las estructuras bronquio izquierdo (LB), bronquio derecho (RB) y carina traqueal (CT) para todos ellos. El resultado de esto es la segmentación anatómica real de 8591 imágenes como *ground truth*, con una contribución media de 245 imágenes por paciente. Finalmente se estableció un *dataset* de las imágenes con sus respectivas etiquetas, que se dividió de acuerdo a una estrategia hold-out de tipo *patient*-wise, resultando en 19 pacientes para el entrenamiento y 16 pacientes para el test, en una proporción 55%-45%. Esta división resulta en un total de 4787 imágenes para el entrenamiento y 4164 imágenes para el grupo de test, de manera que no se han mezclado en ningún momento imágenes de distintos pacientes.

A priori este tamaño muestral parecía insuficiente. Además, existía un desbalanceo de clases en el que la carina traqueal se encontraba subrepresentada con un 30% de presencia relativa frente a una presencia del 36.5% y 33.5% para los bronquios izquierdo y derecho

respectivamente. Gracias a técnicas como *data augmentation* o *probabilistic oversampling* estos problemas no causaron un impacto significativo durante el entrenamiento de la red.

Finalmente, en relación a la población de estudio, si bien los datos recopilados, debido al limitado número de pacientes, no son extrapolables a la población general, nos sirven para inferir las imágenes que nos vamos a encontrar dentro de los pacientes. Un ejemplo puede ser la aparición de sangre en la mucosidad de las vías aéreas. Esto produciría que algunas zonas de interés se hayan visto algo rojizas y se haya podido manchar el broncoscopio tiñendo la imagen. Esto a puede no suponer un problema, ya que la imagen no se debería alterar demasiado, sin embargo, hay determinadas patologías, como las que producen mucosa excesiva en las vías aéreas, que requieren de la eliminación de determinados *frames* en los que se pierden totalmente la imagen y solo se ve mucosidad.

6.2 Modelo de segmentación automática

Para la creación del modelo de segmentación automática, se ha utilizado el *framework* nnU-Net, estado del arte actual en segmentación de imagen médica. Esta herramienta permite una automatización completa de todo el flujo de trabajo y ofrece una gran capacidad de personalización, además de un análisis del dataset que ajusta automáticamente algunos parámetros del entrenamiento. Si bien para el presente TFG se probaron otras herramientas como MONAI, la versatilidad y grado de automatización que ofrece nnU-Net es simplemente superior. El *framework* utilizado no sólo se adapta al *dataset* de estudio, sino que además permite una personalización sobre bloques ya construidos, disminuyendo el tiempo de prototipado y optimizando el flujo de trabajo. MONAI permite arquitecturas más complejas por defecto, y es posible adaptar el entrenamiento al *dataset* de la misma forma que con nnU-Net si se pretendiese, sin embargo, nnU-Net ofrece un mejor compromiso en todos los aspectos además de una presencia y apoyo superiores por la comunidad científica en la literatura.

Durante este trabajo se ha planteado una comparación de dos arquitecturas, una más simple, similar a la U-Net original, denominada Plain U-Net; y otra más compleja, denominada ResEnc U-Net, basada en bloques residuales y muy similar a la red de clasificación ResNet.

Los resultados tras estos entrenamientos nos hacen inclinarnos por la red más simplista, pues es más rápida, menos pesada y con mejores resultados globales. Las métricas de rendimiento como el DSC se inclinan también por ella, al menos en la tarea de segmentación. La arquitectura Plain U-Net supera en *DSC* a la arquitectura ResEnc U-Net para el bronquio derecho y la carina traqueal. Para el bronquio izquierdo este valor es prácticamente igual para ambas arquitecturas, siendo el DSC de la ResEnc U-Net para esta estructura ligeramente superior por 0.0009 puntos. Para otras métricas como el HD95 o *precision* ambas arquitecturas también presentas mucha similitud, siendo la ResEnc U-Net ligeramente superior. No obstante, para las métricas IoU y *recall* la arquitectura Plain U-Net se muestra superior. Finalmente, la arquitectura Plain U-Net supera a la ResEnc U-Net en número de píxeles correctamente clasificados por 103.280 píxeles, algo que, si bien en términos relativos no presenta una gran diferencia, puede ser la causa de un DSC mayor.

6.3 Localización endobronquial

Para la tarea de localización endobronquial también se ha realizado un análisis de sendas arquitecturas, comparando las etiquetas presentes en cada imagen segmentada automáticamente con las etiquetas presentes de cada imagen del *ground truth.* Tras este análisis es posible concluir que la segunda arquitectura, la más compleja, obtiene resultados notablemente mejores que la simple en términos de localización, siendo por tanto más apta para la tarea de navegación. Una red más compleja ha conseguido identificar mejor las estructuras anatómicas, si bien el delineado de estas estructuras es mejorable.

Los artefactos o predicciones “imposibles” suponen un problema significativo a la hora de realizar una localización eficaz. En el análisis de los resultados este problema tiene un efecto pequeño debido a un tamaño muestral reducido, pero que al aumentarlo podría afectar negativamente a la tarea, disminuye la precisión. Este problema sirve para comprobar también con que clases se puede llegar a confundir más la red, demostrando que los grupos LB y RB son los únicos que sufren artefactos, mientras que a la segmentación de la primera bifurcación no le afecta este problema.

6.4 Comparativa con el estado del arte

En relación a estudios similares en este campo, hasta donde conozco, todavía no existen otros trabajos que hayan realizado una segmentación automática de la vía aérea a partir de imágenes de broncoscopia. Es por esto por lo que se plantea una comparación de los modelos creados en el presente TFG frente a modelos de segmentación basados en otro tipo de imágenes, principalmente de TAC, de estructuras dentro del sistema respiratorio, como pulmones, vías aéreas y segmentos pulmonares. Igualmente, se ha considerado comparar, teniendo en cuenta las evidentes diferencias, el rendimiento de los modelos creados frente a modelos de segmentación basados en imagen endoscópica pero no relacionada con las vías aéreas.

En términos de la aplicación de localización, sí que existen estudios que en mayor o menor medida han evaluado la localización de la posición de la sonda del broncoscopio dentro del árbol bronquial a partir de imágenes, lo que facilita la comparación con el presente TFG.

6.4.1 Tarea de segmentación

A continuación, se muestra el análisis y comparación de la red de segmentación de imágenes de broncoscopia frente a redes que segmentan otras estructuras del sistema respiratorio. Posteriormente se realiza este mismo análisis frente a la segmentación de imágenes de endoscopia.

6.4.1.1 Imágenes de broncoscopia frente a otras estructuras pulmonares

En la actual comunidad científica, la métrica de rendimiento más reconocida para caracterizar la calidad de la segmentación de una red es el DSC. En concreto, para la estructura de las **vías aéreas** y su segmentación automática, esta métrica de rendimiento ronda el 0.9 para algunos de los estudios evaluados para estructuras como [77], [78], [79], [80], [81], [82]. El valor de DSC más alto registrado en estos estudios ronda el 0.9495, alcanzado por Wu *et al* [82]. A la vista de estos valores, las redes creadas para la segmentación de imágenes de broncoscopia todavía se encuentran lejos de ser comparables con estos valores. Con un 0.814 de DSC general para la arquitectura Plain

U-Net y un 0.802 para la arquitectura ResEnc U-Net, todavía existe mucho margen de mejora. Si bien las estructuras segmentadas no son exactamente comparables, es preciso destacar que algunos de los anteriores estudios han utilizado también la herramienta nnU-Net, demostrando que es un gran aliado para las tareas de segmentación de imagen.

Frente a la segmentación de la estructura **pulmón** (o **lóbulos pulmonares**, pues en varios estudios se segmentan a la vez) el trabajo realizado sigue la misma tendencia. Debido a la madurez de esta área de estudio, existen trabajos que desarrollan métodos que alcanzan un DSC de 0.988 para la segmentación impulsada por atlas de los lóbulos pulmonares y los pulmones en su conjunto [38], [83], [84]. Es destacable el estudio de Pang *et al*., que haciendo uso de una nnU-Net obtienen valores de DSC de 0.964 [84]. Además, este último estudio también evalúa la métrica HD95, encontrándose entre 4.18 y 7.74 milímetros. Estos valores sí son más comparables con la segmentación propuesta por este trabajo, demostrando que si bien el DSC no logra alcanzar un 0.9, la métrica distancia de Hausdorff resulta bastante prometedora.

Otros métodos también basados en nnU-Net obtienen excelentes resultados (DSC de 0.91) para la segmentación de **vasos sanguíneos pulmonares** [85].

Finalmente, la segmentación de **segmentos pulmonares**, que se trata del área de investigación menos avanzado en este sentido, logra en su actual estado del arte de la mano de Kangxian *et al.* un DSC de 0.8606 empleando funciones implícitas, frente a un 0.8458 empleando nnU-Net.

6.4.1.2 Imágenes de broncoscopia frente a imágenes endoscópicas

En la actualidad no existe una comunidad tan amplia que investigue la segmentación para imágenes endoscópicas como sí existe para imágenes de TC. Sin embargo, existen numerosos estudios que se centran en segmentar elementos representativos de diferentes patologías mediante este tipo de imágenes. La patología que más importancia cobra son los pólipos [86], [87], [88]. Lo más cercano a la segmentación de imágenes de broncoscopia es la segmentación de imágenes de ecografía derivadas de la técnica EBUS durante algunas broncoscopias [89].

En relación a estos estudios, para la segmentación de pólipos, el DSC registrado más alto es de 0.878 (DSC medio obtenido para varios *datasets*), siendo el nuevo estado del arte, frente al 0.768 del anterior mejor resultado. Además, presentan también las métricas de IoU, siendo éstas de 0.809 y 0.702 para el nuevo [86] y el anterior [88] estado del arte, respectivamente.

Analizando estos valores frente a los obtenidos en la segmentación de imágenes de broncoscopia, los resultados alcanzados son mejores que los del anterior estado del arte, pero algo inferiores a los del nuevo. Esto se explica debido a la simplicidad de las arquitecturas que hemos utilizado. El análisis de pólipos utilizando la U-Net convencional ha obtenido un DSC de 0.715, lo que nos indica que es necesario un estudio y modificación de la arquitectura más profundo para lograr rendimiento más elevados.

Frente a las métricas obtenidas de la segmentación de imágenes de ecografía (nodos linfáticos y vasos sanguíneos), con un DSC de 0.71 y 0.76, respectivamente, las redes propuestas mejoran los resultados de la segmentación. Sin embargo, esta comparación no es muy relevante debido al distinto origen de las imágenes.

Analizando estos resultados podemos extraer varias conclusiones. En primer lugar, es importante mencionar que el uso de nnU-Net está muy extendido en la comunidad científica, por lo que utilizar esta herramienta en el estudio le dota de una forma robusta de realizar segmentación de imagen médica. En segundo lugar, es necesario destacar que, si bien debido a la madurez en la segmentación de estructuras como las vías aéreas o los lóbulos pulmonares para imágenes de TC la comparación del DSC no resulta muy beneficiosa para este trabajo, al comparar la propuesta para imágenes de broncoscopia frente a estructuras todavía en investigación la diferencia no es tan notable. En tercer lugar, es relevante que, si comparamos los resultados de la segmentación de imágenes de broncoscopia frente a imágenes de naturaleza similar (imágenes de endoscopia), el rendimiento alcanzado es bastante aceptables. Sin embargo, todavía es necesario optimizar partes de la metodología del estudio para lograr mayores rendimientos predictivos.

En la Tabla 27 se muestra una comparación de algunos de los estudios con mayor rendimiento predictivo en la tarea de segmentación respecto al DSC frente los resultados obtenidos en el presente TFG.

Tabla 27. Tabla comparativa de algunos estudios de gran rendimiento predictivo en estructuras relacionadas con el presente TFG.

Modelo	Autores y fecha	Estructura anatómica	DSC
TfeNet	Wu *et al.* 2025	Vías aéreas	0.950
Atlas-Driven Lung Lobe Segmentation	Zhang *et al.* 2006	Pulmones (lóbulos pulmonares)	0.988
Impulse+	Xie *et al.* 2025	Segmentos pulmonares	0.861
NnU-Net	Mank *et al.* 2025	Vasos sanguíneos pulmonares	0.91
Focus U-Net	Yeung *et al.* 2021	Pólipos durante colonoscopia	0.878
U-Net	Yeung *et al.* 2021	Pólipos durante colonoscopia	0.561
HarDNet-MSEG	Huang *et al.* 2021	Pólipos durante colonoscopia	0.768
Plain U-Net (nnU-Net)	Este TFG. 2025	Imágenes de broncoscopia	0.814
ResEnc U-Net (nnU-Net)	Este TFG. 2025	Imágenes de broncoscopia	0.802

6.4.2 Tarea de localización

Para esta tarea, el principal trabajo al que podemos recurrir para realizar una comparación es el desarrollado por Ying *et al.* [19]. En este estudio se ha creado y validado una red de clasificación de imágenes de broncoscopia con 31 clases distintas a partir de 12 vídeos de broncoscopia distintos. Para las estructuras carina traqueal, bronquio izquierdo y bronquio derecho se han extraído un total de 4161 imágenes (1621 para la carina traqueal, 1426 para el bronquio izquierdo y 1114 para el bronquio derecho).

El tamaño muestral de este estudio es de 342 pacientes para el conjunto de entrenamiento y 12 pacientes para el conjunto de test. Este conjunto es muy superior al utilizado en el presente TFG. Sin embargo, estrategias como el *data augmentation* se han aplicado de forma muy similar a la metodología propuesta, y también se han probado distintas arquitecturas, como se ha hecho también para el trabajo de Ying *et al*.

A nivel de resultados, la métrica utilizada para evaluar el rendimiento predictivo en el estudio de Ying *et al*. es la exactitud o *accuracy*. Esta métrica no se calcula en este trabajo de forma global debido a que el desbalanceo de clases introduce un gran sesgo en la medición para esta medida. Sin embargo, por cada clase, la comparación de los resultados se puede observar en la Tabla 28.

Tabla 28. Comparativa de métricas de accuracy y número de imágenes por estructura con el estado del arte en el contexto de localización.

Zona anatómica	Accuracy ResEnc U-Net	Accuracy Ying *et al.*	Número de imágenes ResEnc U-Net	Número de imágenes Ying *et al.*
Grupo LB	0.91	>0.95	496	1426
Grupo RB	0.85	>0.95	232	1114
Primera bifurcación	0.97	>0.95	3427	1621

Los resultados son comparables para ambos estudios, con una clara ventaja del estudio de Ying *et al.*, que se ha dedicado únicamente a la tarea de clasificación. Debido a que este estudio utiliza más imágenes para el entrenamiento de la red que el TFG propuesto (que claramente tiene una subrepresentación de imágenes de los bronquios derecho e izquierdo estrictamente), al tener resultados comprables se puede valorar positivamente la clasificación mediante imágenes segmentadas. Esto nos indica que sí es posible realizar una localización eficaz basada únicamente en imágenes de segmentación y sin la necesidad de un *dataset* tan extenso, aunque esto mejoraría sin duda los resultados.

6.5 Limitaciones

Una vez finalizado el estudio y habiéndolo comparado con las propuestas más prometedoras de diversas áreas de segmentación y localización de imagen médica, se proceden a destacar las principales limitaciones de este TFG, así como sus posibilidades de mejora.

En relación a los vídeos de broncoscopia, una limitación que puede haber afectado al desempeño de la red es la presencia del marco negro e información complementaria en la pantalla durante las broncoscopias con el equipo OLYMPUS. Esto ha supuesto un reto para la segmentación manual, así como una posible pérdida de poder de segmentación.

Relacionado con lo anterior, al inicio del estudio se planteó una segmentación automática de todas las estructuras del árbol bronquial. Debido a numerosos factores como el carácter prospectivo del estudio, el tiempo necesario para realizar una segmentación completa de un paciente y el tiempo disponible para llevar esto a cabo, se optó por limitar el estudio a las tres estructuras más características del árbol bronquial: la primera bifurcación y los bronquios derivados de esta. La causa que esto conlleva es

menos imágenes totales, pero más localizadas. Adicionalmente, se plantea como una limitación la curva de aprendizaje de la segmentación manual, puesto que las últimas segmentaciones son de mayor calidad que las iniciales.

Además de esto, al principio del estudio también se pretendía alcanzar un mayor tamaño muestral. Con un tiempo de recopilación tan limitado esto ha sido imposible, dando lugar a un *dataset* pequeño y poco robusto. Es por esto por lo que ha sido necesario un enfoque muy agresivo de *data augmentation* y un gran número de iteraciones por época para subsanar esta falta de pacientes.

En relación a la etapa de aprendizaje, una gran limitación ha sido el tiempo de entrenamiento. Con una herramienta como nnU-Net, este tiempo tan extenso se ha compensado con una curva de aprendizaje relativamente sencilla y una gran facilidad de uso. Sin embargo, con la finalidad de mejorar los resultados debido a la falta de pacientes, el número de segundos por época ha aumentado drásticamente, teniendo que estar días enteros entrenando modelos para comprobar los resultados intermedios y evaluar opciones de mejora. Adicionalmente, debido a que nnU-Net durante el entrenamiento ofrece pseudométricas basadas en parches, en muchas ocasiones éstas no se corresponden con la realidad y es necesario terminar un entrenamiento de 17 horas para realizar una evaluación global del modelo para verificar si efectivamente la tendencia se mantenía.

Para compensar también en parte estos tiempos de entrenamiento tan elevados, se redujo el número de épocas. Esto produce resultados menos robustos, aunque en general se ha demostrado que es un buen compromiso tiempo-robustez.

Por último, en relación a la funcionalidad de localización, identificar una zona anatómica específica a partir de imágenes segmentadas podría pensarse que es una forma poco precisa de localizarse dentro de un espacio tan especializado. Sin embargo, a pesar de tratarse de un enfoque bastante generalista, se ha demostrado que ofrece buenos resultados cuando la intención de uso es una localización rápida y precisa de estructuras del árbol bronquial, más específicamente el bronquio derecho, el bronquio izquierdo o la carina traqueal.

7. Conclusiones

La broncoscopia es una técnica endoscópica muy utilizada durante la práctica clínica habitual en un Servicio de Neumología, que permite visualizar las vías del sistema respiratorio y acceder físicamente al árbol bronquial con numerosas posibilidades diagnósticas y terapéuticas. En la actualidad, existen algunas formas de navegación endobronquial centradas principalmente en la imagenología complementaria, como radiografías a pie de cama o ecografías. Sin embargo, las nuevas tendencias como la cirugía robótica o la inteligencia artificial están ejerciendo un efecto en esta área y están naciendo nuevas técnicas para este cometido.

Este trabajo nace de la necesidad de realizar una localización precisa dentro del árbol bronquial, por lo que se ha propuesto realizar esta tarea a partir de imágenes de broncoscopia segmentadas.

Para crear un proceso de segmentación automática a partir de los datos recopilados, se ha utilizado la herramienta nnU-Net. Este *framework* permite la automatización casi total del procedimiento de preprocesado, entrenamiento, evaluación y postprocesado, siendo el estado del arte en la segmentación de numerosas estructuras anatómicas. Gracias a este, se han analizado y comparado dos arquitecturas de red distintas: Plain U-Net y ResEnc U-Net. La diferencia principal de estas arquitecturas reside en que mientras que la primera se muestra como una adaptación simple de la arquitectura U-Net original, la segunda utiliza el concepto de conexiones residuales para manejar problemas en la propagación del gradiente.

Tras el entrenamiento, se han evaluado las métricas de rendimiento de los modelos, siendo la métrica más relevante el DSC (*Dice Similarity Coefficient*). Plain U-Net ha obtenido un DSC de 0.814 y ResEnc U-Net de 0.802, demostrando que no es necesaria una arquitectura más compleja para obtener rendimientos predictivos elevados.

A partir de la creación de estas redes se demuestra también que no es necesario un postprocesado, puesto que las redes entrenadas tienen la robustez suficiente como para no crear falsos positivos aislados o presentar unos bordes demasiado irregulares.

Finalmente, llegados a la tarea de localización, se concluye que es posible utilizar imágenes de broncoscopia segmentadas automáticamente para realizar una localización bronquial. A partir de ésta se obtienen como resultado métricas de rendimiento elevadas derivadas de la matriz de confusión multiclase (PPV, *recall*, *F1 – score* y *accuracy*) para ambas arquitecturas, que distinguen en localización entre la primera bifurcación, el bronquio izquierdo y el bronquio derecho. Con esta localización se consigue demostrar también que la red con mejor DSC no es la que mejores resultados de localización presenta, pues la red ResEnc U-Net termina siendo más efectiva para esta tarea que Plain U-Net.

7.1 Contribuciones a la innovación y al estado del arte

El procesado automático de imagen médica es un tema muy presente en la Neumología actual. Hasta donde conocemos, se han segmentado imágenes de resonancia magnética y tomografía computacional, pero no de broncoscopia. Lo mismo sucede con la imagenología endoscópica general, pues tradicionalmente no es un tipo de imagen demasiado estudiado por la comunidad científica. Sin embargo, la segmentación de imágenes de broncoscopia, si bien ha permanecido como algo inexplorado, puede resultar de utilidad para diversas aplicaciones, como así ha sido demostrado.

Para ello, lo primero y más fundamental ha sido crear un *dataset* de estas imágenes, ya que no existen repositorios públicos que las suministren. Durante el proceso de recopilación de los vídeos de broncoscopia en el Hospital Universitario Río Hortega, se ha procedido a segmentar manualmente la carina traqueal, el bronquio izquierdo y el bronquio derecho de todos ellos. Finalmente, se ha cerrado la etapa de recopilación con 35 vídeos de pacientes distintos y un total de 8591 imágenes segmentadas, divididas en entrenamiento y test en una proporción de 55%-45%.

Es por esto por lo que la mayor contribución de este trabajo ha sido la exploración de un tipo de imagen cuyo procesado está todavía dando sus primeros pasos. Las aportaciones que esta exploración ha supuesto a la investigación de imágenes de broncoscopia son:

1. Creación del primer *dataset* de imágenes de broncoscopia segmentadas.

2. Creación de la primera red de segmentación automática para imágenes de broncoscopia a nivel de primera bifurcación y bronquios derecho e izquierdo.
3. Creación de un método de localización endobronquial a partir de imágenes segmentadas de broncoscopia.

A estas tres contribuciones le podemos sumar la realización de un estado del arte de un área de la ciencia en constante evolución, que puede servir como apoyo a investigadores que se adentren en esta área.

7.2 Conclusiones principales del estudio

A continuación, se enumeran las principales conclusiones extraídas tras la realización del TFG:

1. El procesado y almacenamiento de imágenes de broncoscopia como imágenes adaptadas a una estructura de tipo TC resultan de utilidad no solo para crear un *dataset* compacto y ordenado, sino también para mejorar el rendimiento del entrenamiento de las redes que utilicen estos datos.
2. Es posible segmentar automáticamente las estructuras principales del árbol bronquial. Además, no es necesario utilizar una arquitectura demasiado compleja para lograr rendimientos predictivos elevados.
3. Es posible realizar una localización bronquial a partir de imágenes de broncoscopia segmentadas automáticamente. Para esta tarea se ha comprobado que, cuanto mayor sea la complejidad de la red, mejores resultados se obtendrán y mayor capacidad de detección tendrá la herramienta.
4. El postprocesado para las imágenes segmentadas automáticamente basado en componentes conectados no es necesario, debido a que las redes tienen la robustez suficiente como para no delinear falsos positivos aislados a lo largo de la imagen.
5. El postprocesado para las imágenes segmentadas automáticamente basado en el suavizado gaussiano empeora algunas métricas de validación, como el DSC o el IoU, ya que modifica los bordes de las segmentaciones, volviéndolos más lisos, pero alterando significativamente el contorno de las segmentaciones y aumentando tanto el número de verdaderos como de falsos positivos.

Finalmente, después de analizar y comparar los resultados obtenidos en este TFG frente a los resultados de trabajos similares y teniendo en cuenta las diferencias metodológicas que en mayor o menor medida condicionan esta comparativa, se puede

concluir que, si bien la segmentación de imágenes de broncoscopia es un área poco explorada, el trabajo realizado cumple con los estándares de DSC aceptables y la localización en el árbol bronquial a partir de imágenes segmentadas es comparable a la clasificación de imágenes sin segmentar obteniendo unos rendimientos predictivos similares.

7.3 Líneas futuras de investigación

De acuerdo a las conclusiones anteriores, se plantean ahora una serie de posibles líneas de investigación con la intención de mejorar el estudio realizado y establecer como una alternativa clínica viable a las técnicas actuales la localización bronquial mediante imágenes segmentadas automáticamente.

La primera línea de investigación giraría en torno a ampliar el tiempo de recopilación de vídeos y a hacerlo de manera multicéntrica. Puesto que el estudio se ha realizado en un único centro, para dos equipos distintos, estos resultados pueden no ser extrapolables para otros centros con otros equipamientos. Es por esto por lo que se podría plantear un estudio que comprenda a varios hospitales de Castilla y León o incluso a nivel de toda España.

Unos resultados prometedores con 35 pacientes pueden convertirse en una opción sólida de la localización bronquial con 500 o 1000 pacientes de distintos centros y con imágenes tomadas desde distintos equipos. Para ello, se debería aumentar el número de imágenes por paciente, el número de pacientes, y el tiempo de recopilación. También sería interesante crear una base de datos pública de estas imágenes segmentadas, con el objetivo de contribuir a la ciencia abierta.

Respecto a este punto, también es posible ampliar el estudio a la detección de anomalías de la anatomía del árbol bronquial, como pueden ser patologías como la hemoptisis o la presencia de cuerpos extraños; o simplemente variantes anatómicas difíciles de detectar por los profesionales.

La segunda línea de investigación se centraría en utilizar las redes creadas durante este estudio como herramientas base para realizar fine-tuning con otras redes de segmentación de índole endoscópica. Con esta estrategia ya se ha demostrado que es posible mejorar los resultados si ambos problemas se parecen lo suficiente. Debido al gran parecido entre las imágenes broncoscópicas y endoscópicas, sería posible mejorar

los resultados de las investigaciones de imagen endoscópica. Además, para robótica quirúrgica sería posible también crear un navegador endoscópico basado en imágenes endoscópicas segmentadas automáticamente. Realizando inferencias en tiempo real, esta estrategia podría asentar también las bases de la cirugía autónoma, puesto que el robot quirúrgico tomaría decisiones en función de lo que está viendo y segmentando en directo.

Finalmente, una tercera línea futura simplemente podría centrarse en la mejora de etapas relevantes o críticas dentro de la metodología de este estudio. Partiendo del *dataset raw*, se podría probar la influencia de distintos preprocesados o normalizaciones sobre los datos para maximizar el rendimiento predictivo de la red. Se pueden alterar también los hiperparámetros para compensar de una forma más eficiente la falta de datos, y también se pueden adaptar arquitecturas de U-net más novedosas para lograr mejores resultados. Por último, sería interesante valorar el rendimiento de los *vision transformers* (ViT) para esta tarea, ya que esta opción no ha sido contemplada en este estudio y según las tendencias actuales de la inteligencia artificial estas nuevas arquitecturas se postulan como el estado del arte en numerosas tareas de visión artificial.

8. Bibliografía

[1] T. S. Panchabhai, M. Ghobrial, and A. C. Mehta, “History of Bronchoscopy: The Evolution of Interventional Pulmonology,” in *Interventions in Pulmonary Medicine*, Cham: Springer International Publishing, 2018, pp. 609–621. doi: 10.1007/978-3-319-58036-4_39.

[2] G. J. Criner *et al.*, “Interventional Bronchoscopy,” *Am J Respir Crit Care Med*, vol. 202, no. 1, pp. 29–50, Jul. 2020, doi: 10.1164/rccm.201907-1292SO.

[3] T. Ishiwata, A. Gregor, T. Inage, and K. Yasufuku, “Bronchoscopic navigation and tissue diagnosis,” *Gen Thorac Cardiovasc Surg*, vol. 68, no. 7, pp. 672–678, Jul. 2020, doi: 10.1007/s11748-019-01241-0.

[4] M. Shafiq, H. Lee, L. Yarmus, and D. Feller-Kopman, “Recent Advances in Interventional Pulmonology,” *Ann Am Thorac Soc*, vol. 16, no. 7, pp. 786–796, Jul. 2019, doi: 10.1513/AnnalsATS.201901-044CME.

[5] Y. B. Gesthalter and C. L. Channick, “Interventional Pulmonology: Extending the Breadth of Thoracic Care,” *Annu Rev Med*, vol. 75, no. 1, pp. 263–276, Jan. 2024, doi: 10.1146/annurev-med-050922-060929.

[6] G. Michaud, “Review of Recent Important Papers in Interventional Pulmonology,” *Semin Thorac Cardiovasc Surg*, vol. 30, no. 2, pp. 212–214, 2018, doi: https://doi.org/10.1053/j.semtcvs.2018.05.003.

[7] F. Xie, A. Wagh, R. Wu, D. K. Hogarth, and J. Sun, “Robotic-assisted bronchoscopy in the diagnosis of peripheral pulmonary lesions,” *Chinese Medical Journal Pulmonary and Critical Care Medicine*, vol. 1, no. 1, pp. 30–35, Mar. 2023, doi: 10.1016/j.pccm.2023.01.001.

[8] L. Zhu, J. Zheng, C. Wang, J. Jiang, and A. Song, “A bronchoscopic navigation method based on neural radiation fields,” *Int J Comput Assist Radiol Surg*, vol. 19, no. 10, pp. 2011–2021, Aug. 2024, doi: 10.1007/s11548-024-03243-7.

[9] “Bronchoskopie - Elektromagnetisch navigierte Bronchoskopie steigert Erfolgsrate,” *Pneumologie*, vol. 61, no. 11, pp. 688–688, Nov. 2007, doi: 10.1055/s-2007-998821.

[10] J. Collins, P. Goldstraw, and Paul. Dhillon, *Practical bronchoscopy*. Oxford ; Boston: Blackwell Scientific Publications ; Chicago, Ill. : Year Book Medical Publishers [distributor], 1987.

[11] L. Vaz Rodrigues, Y. Martins, C. Guimarães, M. de Santis, A. Marques, and F. Barata, “Anatomy for the bronchologist: A prospective study of the normal endobronchial anatomic variants,” *Revista Portuguesa de Pneumologia (English Edition)*, vol. 17, no. 5, pp. 211–215, Sep. 2011, doi: 10.1016/j.rppnen.2011.06.004.

[12] C. Di Felice, P. Ntiamoah, and T. R. Gildea, “Bronchoscopic Anatomy,” in *Video-Atlas of VATS Pulmonary Sublobar Resections*, Cham: Springer International Publishing, 2023, pp. 25–31. doi: 10.1007/978-3-031-14455-4_3.

[13] Z. Zhang, S. Li, and Y. Bao, “Endobronchial Ultrasound-Guided Transbronchial Mediastinal Cryobiopsy versus Endobronchial Ultrasound-Guided Transbronchial Needle

Aspiration for Mediastinal Disorders: A Meta-Analysis," *Respiration*, vol. 103, no. 7, pp. 359–367, 2024, doi: 10.1159/000538609.

[14] M. Terranova Ríos *et al.*, "Endobronchial Ultrasound-Guided Transbronchial Mediastinal Cryobiopsy: Tunneling With Different Needle Gauges," *Open Respiratory Archives*, vol. 7, no. 2, p. 100405, Apr. 2025, doi: 10.1016/j.opresp.2025.100405.

[15] A. T. C. Chang, J. W. Y. Chan, I. C. H. Siu, W. Liu, R. W. H. Lau, and C. S. H. Ng, "Robotic-assisted bronchoscopy—advancing lung cancer management," *Front Surg*, vol. 12, Jun. 2025, doi: 10.3389/fsurg.2025.1566902.

[16] S. Fernandez-Bussy *et al.*, "Robotic-assisted bronchoscopy: a narrative review of systems," *J Thorac Dis*, vol. 16, no. 8, pp. 5422–5434, Aug. 2024, doi: 10.21037/jtd-24-456.

[17] Q. Tian *et al.*, "DD-VNB: A Depth-based Dual-Loop Framework for Real-time Visually Navigated Bronchoscopy," Mar. 2024, doi: 10.1109/IROS58592.2024.10801553.

[18] Q. Tian *et al.*, "PANS: Probabilistic Airway Navigation System for Real-time Robust Bronchoscope Localization," Jul. 2024.

[19] Y. Li *et al.*, "Development and validation of the artificial intelligence (AI)-based diagnostic model for bronchial lumen identification," *Transl Lung Cancer Res*, vol. 11, no. 11, pp. 2261–2274, Nov. 2022, doi: 10.21037/tlcr-22-761.

[20] Q. Tian *et al.*, "BronchoTrack: Airway Lumen Tracking for Branch-Level Bronchoscopic Localization," Feb. 2024.

[21] E. Ramírez, C. Sánchez, A. Borràs, M. Diez-Ferrer, A. Rosell, and D. Gil, "BronchoX: bronchoscopy exploration software for biopsy intervention planning," *Healthc Technol Lett*, vol. 5, no. 5, pp. 177–182, Oct. 2018, doi: 10.1049/htl.2018.5074.

[22] S. F. Ashraf and K. K. W. Lau, "Navigation bronchoscopy: A new tool for pulmonary infections.," *Med Mycol*, vol. 57, no. Supplement_3, pp. S287–S293, Jun. 2019, doi: 10.1093/mmy/myz058.

[23] J. S. Wang Memoli, P. J. Nietert, and G. A. Silvestri, "Meta-analysis of Guided Bronchoscopy for the Evaluation of the Pulmonary Nodule," *Chest*, vol. 142, no. 2, pp. 385–393, Aug. 2012, doi: 10.1378/chest.11-1764.

[24] Annett Zündorf, "Bronchoskopie - Elektromagnetisch navigierte Bronchoskopie steigert Erfolgsrate," *Pneumologie*, vol. 61, no. 11, pp. 688–688, Nov. 2007, doi: 10.1055/s-2007-998821.

[25] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Comput Sci*, vol. 2, no. 6, p. 420, Nov. 2021, doi: 10.1007/s42979-021-00815-1.

[26] S. Haykin, *Neural Networks: A Comprehensive Foundation. 2nd Edition. Prentice Hall International, Upper Saddle River.*, 2nd Edition. Hamilton, Ontario, Canada: Pearson Education, 1999.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[28] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Med Image Anal*, vol. 89, p. 102918, Oct. 2023, doi: 10.1016/J.MEDIA.2023.102918.

[29] Y. Yu *et al.*, "Techniques and Challenges of Image Segmentation: A Review," *Electronics 2023, Vol. 12, Page 1199*, vol. 12, no. 5, p. 1199, Mar. 2023, doi: 10.3390/ELECTRONICS12051199.

[30] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int J Comput Vis*, vol. 1, no. 4, pp. 321–331, Jan. 1988, doi: 10.1007/BF00133570.

[31] S. A. Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep Semantic Segmentation of Natural and Medical Images: A Review," Mar. 2024.

[32] S. Ghosh, N. Das, I. Das, and U. Maulik, "Understanding Deep Learning Techniques for Image Segmentation," *ACM Comput Surv*, vol. 52, no. 4, pp. 1–35, Jul. 2020, doi: 10.1145/3329784.

[33] Z. Tian, B. Zhang, H. Chen, and C. Shen, "Instance and Panoptic Segmentation Using Conditional Convolutions," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 1, pp. 669–680, Jan. 2023, doi: 10.1109/TPAMI.2022.3145407.

[34] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *IEEE Trans Pattern Anal Mach Intell*, pp. 1–1, 2021, doi: 10.1109/TPAMI.2021.3059968.

[35] K. Kuang *et al.*, "What Makes for Automatic Reconstruction of Pulmonary Segments," Jul. 2022.

[36] K. Xie *et al.*, "Template-Guided Reconstruction of Pulmonary Segments with Neural Implicit Functions," May 2025.

[37] Z. Pápai-Székely, G. Grmela, and V. Sárosi, "Novel diagnostic processes and challenges in bronchoscopy," *Pathology and Oncology Research*, vol. 30, May 2024, doi: 10.3389/pore.2024.1611774.

[38] E. M. van Rikxoort and B. van Ginneken, "Automated segmentation of pulmonary structures in thoracic computed tomography scans: a review," *Phys Med Biol*, vol. 58, no. 17, pp. R187–R220, Sep. 2013, doi: 10.1088/0031-9155/58/17/R187.

[39] Z. Chen *et al.*, "Deep learning-based bronchial tree-guided semi-automatic segmentation of pulmonary segments in computed tomography images," *Quant Imaging Med Surg*, vol. 14, no. 2, pp. 1636–1651, Feb. 2024, doi: 10.21037/qims-23-1251.

[40] S. G. Armato and W. F. Sensakovic, "Automated lung segmentation for thoracic CT," *Acad Radiol*, vol. 11, no. 9, pp. 1011–1021, Sep. 2004, doi: 10.1016/j.acra.2004.06.005.

[41] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nat Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021, doi: 10.1038/s41592-020-01008-z.

[42] F. Isensee *et al.*, "nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation," Apr. 2024, doi: 10.1007/978-3-031-72114-4_47.

[43] C.-M. Hsu, C.-C. Hsu, Z.-M. Hsu, F.-Y. Shih, M.-L. Chang, and T.-H. Chen, "Colorectal Polyp Image Detection and Classification through Grayscale Images and Deep Learning," *Sensors*, vol. 21, no. 18, p. 5995, Sep. 2021, doi: 10.3390/s21185995.

[44] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, “Segment anything in medical images,” *Nat Commun*, vol. 15, no. 1, p. 654, Jan. 2024, doi: 10.1038/s41467-024-44824-z.

[45] Y. Süküt, E. Yurdakurban, and G. S. Duran, “Accuracy of deep learning-based upper airway segmentation,” *J Stomatol Oral Maxillofac Surg*, vol. 126, no. 2, p. 102048, Mar. 2025, doi: 10.1016/j.jormas.2024.102048.

[46] A. Lo Giudice, V. Ronsivalle, G. Gastaldi, and R. Leonardi, “Assessment of the accuracy of imaging software for 3D rendering of the upper airway, usable in orthodontic and craniofacial clinical settings,” *Prog Orthod*, vol. 23, no. 1, p. 22, Dec. 2022, doi: 10.1186/s40510-022-00413-8.

[47] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[48] H. Taud and J. F. Mas, “Multilayer Perceptron (MLP),” 2018, pp. 451–455. doi: 10.1007/978-3-319-60801-3_27.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[50] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, “Medical Image Analysis using Convolutional Neural Networks: A Review,” *J Med Syst*, vol. 42, no. 11, p. 226, Nov. 2018, doi: 10.1007/s10916-018-1088-1.

[51] L. Madhuanand, P. Sadavarte, A. J. H. Visschedijk, H. A. C. Denier Van Der Gon, I. Aben, and F. B. Osei, “Deep convolutional neural networks for surface coal mines determination from sentinel-2 images,” *Eur J Remote Sens*, vol. 54, no. 1, pp. 296–309, Jan. 2021, doi: 10.1080/22797254.2021.1920341.

[52] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Apr. 2015.

[53] N. S. Punn and S. Agarwal, “Modality specific U-Net variants for biomedical image segmentation: a survey,” *Artif Intell Rev*, vol. 55, no. 7, pp. 5845–5889, Oct. 2022, doi: 10.1007/s10462-022-10152-1.

[54] F. Isensee and K. H. Maier-Hein, “An attempt at beating the 3D U-Net,” Oct. 2019.

[55] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[56] A. Myronenko, D. Yang, Y. He, and D. Xu, “Automated 3D Segmentation of Kidneys and Tumors in MICCAI KiTS 2023 Challenge,” Oct. 2023.

[57] R. Nawarathna *et al.*, “Abnormal image detection in endoscopy videos using a filter bank and local binary patterns,” *Neurocomputing*, vol. 144, pp. 70–91, Nov. 2014, doi: 10.1016/j.neucom.2014.02.064.

[58] J. Hamwood, D. Alonso-Caneiro, S. A. Read, S. J. Vincent, and M. J. Collins, “Effect of patch size and network architecture on a convolutional neural network approach for automatic segmentation of OCT retinal layers.,” *Biomed Opt Express*, vol. 9, no. 7, pp. 3049–3066, Jul. 2018, doi: 10.1364/BOE.9.003049.

[59] G. I. Quintana, Z. Li, L. Vancamberg, M. Mougeot, A. Desolneux, and S. Muller, "Exploiting Patch Sizes and Resolutions for Multi-Scale Deep Learning in Mammogram Image Classification," *Bioengineering*, vol. 10, no. 5, p. 534, Apr. 2023, doi: 10.3390/bioengineering10050534.

[60] J. Sato and S. Kido, "Large Batch and Patch Size Training for Medical Image Segmentation," Oct. 2022.

[61] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, Sep. 1951, doi: 10.1214/aoms/1177729586.

[62] J. Kiefer and J. Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462–466, Sep. 1952, doi: 10.1214/aoms/1177729392.

[63] Y. Nesterov, "A method for solving the convex programming problem with convergence rate O(1/k^2)," *Proceedings of the USSR Academy of Sciences*, vol. 269, pp. 543–547, Jan. 1983.

[64] X. Xie, P. Zhou, H. Li, Z. Lin, and S. Yan, "Adan: Adaptive Nesterov Momentum Algorithm for Faster Optimizing Deep Models," Aug. 2022.

[65] R. Azad *et al.*, "Loss Functions in the Era of Semantic Segmentation: A Survey and Outlook," Dec. 2023.

[66] A. Galdran, G. Carneiro, and M. Á. G. Ballester, "On the Optimal Combination of Cross-Entropy and Soft Dice Losses for Lesion Segmentation with Out-of-Distribution Robustness," Sep. 2022.

[67] D. Capellán-Martín *et al.*, "Model Ensemble for Brain Tumor Segmentation in Magnetic Resonance Imaging," Sep. 2024, doi: 10.1007/978-3-031-76163-8_20.

[68] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance Normalization: The Missing Ingredient for Fast Stylization," Nov. 2017.

[69] A. L. Maas, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:16489696

[70] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a Guideline for Evaluation Metrics in Medical Image Segmentation," Feb. 2022.

[71] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.

[72] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Med Imaging*, vol. 15, no. 1, p. 29, Dec. 2015, doi: 10.1186/s12880-015-0068-x.

[73] X. Chen *et al.*, "CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy," *Front Oncol*, vol. 10, Apr. 2020, doi: 10.3389/fonc.2020.00524.

[74] P. Bilic *et al.*, "The Liver Tumor Segmentation Benchmark (LiTS)," Nov. 2022, doi: 10.1016/j.media.2022.102680.

[75] N. Heller *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge," *Med Image Anal*, vol. 67, p. 101821, Jan. 2021, doi: 10.1016/j.media.2020.101821.

[76] H. Lin *et al.*, "Gaussian filter facilitated deep learning-based architecture for accurate and efficient liver tumor segmentation for radiation therapy," *Front Oncol*, vol. 14, Jun. 2024, doi: 10.3389/fonc.2024.1423774.

[77] A. Garcia-Uceda, R. Selvan, Z. Saghir, H. A. W. M. Tiddens, and M. de Bruijne, "Automatic airway segmentation from computed tomography using robust and efficient 3-D convolutional neural networks," *Sci Rep*, vol. 11, no. 1, p. 16001, Aug. 2021, doi: 10.1038/s41598-021-95364-1.

[78] Y. Yuan *et al.*, "An end-to-end multi-scale airway segmentation framework based on pulmonary CT image," *Phys Med Biol*, vol. 69, no. 11, p. 115027, Jun. 2024, doi: 10.1088/1361-6560/ad4300.

[79] S. Yuan, "Extraction of Pulmonary Airway in CT Scans Using Deep Fully Convolutional Networks," Aug. 2022.

[80] Y. Süküt, E. Yurdakurban, and G. S. Duran, "Accuracy of deep learning-based upper airway segmentation," *J Stomatol Oral Maxillofac Surg*, vol. 126, no. 2, p. 102048, Mar. 2025, doi: 10.1016/j.jormas.2024.102048.

[81] B. Yang *et al.*, "Progressive Curriculum Learning with Scale-Enhanced U-Net for Continuous Airway Segmentation," Feb. 2025.

[82] Q. Wu, Y. Wang, and Q. Zhang, "Airway Segmentation Network for Enhanced Tubular Feature Extraction," Jul. 2025.

[83] Li Zhang, E. A. Hoffman, and J. M. Reinhardt, "Atlas-driven lung lobe segmentation in volumetric X-ray CT images," *IEEE Trans Med Imaging*, vol. 25, no. 1, pp. 1–16, Jan. 2006, doi: 10.1109/TMI.2005.859209.

[84] H. Pang *et al.*, "A fully automatic segmentation pipeline of pulmonary lobes before and after lobectomy from computed tomography images," *Comput Biol Med*, vol. 147, p. 105792, Aug. 2022, doi: 10.1016/j.compbiomed.2022.105792.

[85] Q. J. Mank *et al.*, "Artificial intelligence-based pulmonary vessel segmentation: an opportunity for automated three-dimensional planning of lung segmentectomy.," *Interdisciplinary cardiovascular and thoracic surgery*, vol. 40, no. 5, May 2025, doi: 10.1093/icvts/ivaf101.

[86] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, "Focus U-Net: A novel dual attention-gated CNN for polyp segmentation during colonoscopy," Jun. 2021.

[87] T. Mahmud, B. Paul, and S. A. Fattah, "PolypSegNet: A modified encoder-decoder architecture for automated polyp segmentation from colonoscopy images," *Comput Biol Med*, vol. 128, p. 104119, Jan. 2021, doi: 10.1016/j.compbiomed.2020.104119.

[88] C.-H. Huang, H.-Y. Wu, and Y.-L. Lin, "HarDNet-MSEG: A Simple Encoder-Decoder Polyp Segmentation Neural Network that Achieves over 0.9 Mean Dice and 86 FPS," Jan. 2021.

[89] Ø. Ervik *et al.*, "Automatic Segmentation of Mediastinal Lymph Nodes and Blood Vessels in Endobronchial Ultrasound (EBUS) Images Using Deep Learning," *J Imaging*, vol. 10, no. 8, p. 190, Aug. 2024, doi: 10.3390/jimaging10080190.