DEPARTAMENTO DE
INGENIERÍA ELÉCTRICA
UNIVERSIDAD DE CONCEPCIÓN

Segmentación automática de lesiones cutáneas usando deep-learning

POR

Benjamin Ismael Aguayo Mellado

Memoria de Título presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al título profesional de Ingeniero Civil en Telecomunicaciones.

Profesor guia: Sebastián Godoy Medel
Sergio Torres Inostroza
Francisco Pérez Venegas

Concepción,
23 de enero de 2026

A mi familia, por su apoyo incondicional durante estos años de universidad.

Resumen

En la practica clínica, las imágenes dermatologicas suelen capturarse bajo condiciones heterogeneas de iluminación, enfoque y calidad, lo que puede afectar el desempeño de sistemas automáticos basados en aprendizaje profundo. Este trabajo tiene como objetivo evaluar la confiabilidad de la segmentación automática de lesiones cutáneas cuando la calidad de imagen se degrada, identificando condiciones limite donde la segmentación deja de ser operable fuera de entornos clínicos ideales. Para ello se implementan y comparan tres arquitecturas de segmentación (U-Net, DeepLabV3+ y Mask R-CNN) mediante un modelo metodológico de evaluación basado en un sistema modular de siete bloques de degradación, que permite generar perturbaciones controladas y describir el comportamiento del sistema bajo condiciones de estrés en la adquisición. La robustez se valida con Intersección sobre Unión (IoU), coeficiente Dice, sensibilidad (recall) y precisión, utilizando los conjuntos de datos ISIC 2018 (International Skin Imaging Collaboration 2018) y PH2, y se adopta un criterio operacional (IoU >= 0.75, Dice >= 0.85, recall >= 0.85, precision >= 0.80) para determinar limites de uso confiable.

En condicion base, U-Net alcanza IoU 0.8284 y Dice 0.893. La reducción de resolución mantiene en general el recall por sobre 85 %, aunque deteriora el contorno en niveles extremos. El factor mas critico es el ruido gaussiano: DeepLabV3+ reduce su recall hasta 51.6 % (varianza 0.03), mientras U-Net conserva recall alto pero disminuye su precision a 77.7 %, evidenciando sobre segmentacion. Mask R-CNN presenta mayor estabilidad frente a ruido, manteniendo recall sobre 87 % y precision sobre 83 % en el peor caso evaluado. Estos resultados delimitan rangos de calidad de imagen donde cada arquitectura mantiene un desempeno consistente y muestran que el ruido gaussiano severo condiciona la operabilidad del sistema.

Abstract

In clinical practice, dermatological images are often acquired under heterogeneous lighting, focus, and overall quality conditions, which can compromise the performance of deep learning based automated systems. This work aims to evaluate the reliability of automatic skin lesion segmentation as image quality degrades, identifying boundary conditions where segmentation becomes non operational outside ideal settings. To this end, three segmentation architectures (U-Net, DeepLabV3+, and Mask R-CNN) are implemented and compared using a formal evaluation model based on a modular system of seven degradation blocks, which enables controlled perturbations and characterizes system behavior under acquisition stress conditions. Robustness is assessed using Intersection over Union (IoU), Dice coefficient, sensitivity (recall), and precision on the ISIC 2018 (International Skin Imaging Collaboration 2018) and PH2 datasets, and an operational criterion (IoU >= 0.75, Dice >= 0.85, recall >= 0.85, precision >= 0.80) is adopted to determine reliable use limits.

Under baseline conditions, U-Net achieves IoU 0.8284 and Dice 0.893. Resolution reduction generally keeps recall above 85 %, but degrades boundary delineation at extreme levels. The most critical factor is Gaussian noise: DeepLabV3+ reduces recall down to 51.6 % (variance 0.03), while U-Net maintains high recall but drops to 77.7 % precision, indicating over segmentation. Mask R-CNN shows higher stability under noise, maintaining recall above 87 % and precision above 83 % in the worst evaluated case. These results delineate image quality ranges in which each architecture maintains consistent performance and show that severe Gaussian noise conditions constrain system operability. . .

Agradecimientos

Mi gratitud se dirige a la Universidad de Concepción y, de manera especial, al Departamento de Ingeniería Eléctrica, por proporcionar el rigor académico, los recursos y el ambiente de innovación esenciales para mi desarrollo como Ingeniero en Telecomunicaciones.

Agradezco sinceramente a mi profesor guía, el Dr. Sebastián Eugenio Godoy Medel, por su invaluable orientación, su paciencia y su constante apoyo a lo largo del desarrollo de esta investigación.

Finalmente, y de manera muy especial, agradezco a mi familia, por su amor incondicional, su comprensión y su estímulo permanente. A mis amigos, por el apoyo incondicional y por ser un pilar de motivación durante esta etapa. A todos ellos, por su confianza.

Índice General

Índice de Figuras

Índice de Tablas

Siglas

ASPP Atrous Spatial Pyramid Pooling

CAD Diagnóstico Asistido por Computadora

CNN Red Neuronal Convolucional

COCO Common Objects in Context

Dice Dice Coefficient

DL Aprendizaje Profundo

DSC Dice Similarity Coefficient

FAIR Facebook AI Research

FN Falso Negativo

FP Falso Positivo

FPN Feature Pyramid Network

IoU Intersección sobre Unión

ISIC International Skin Imaging Collaboration

R-CNN Region-based Convolutional Neural Network

ReLU Rectified Linear Unit

ResNet Residual Neural Network

s&p Ruido Sal y Pimienta

TP Verdadero Positivo

Capítulo 1

Introducción

La detección temprana de lesiones cutáneas, en particular aquellas con potencial maligno, es un desafió relevante para la salud publica debido a que un diagnostico oportuno influye directamente en el pronostico del paciente. En la practica clínica, la evaluación inicial se apoya en la inspección visual y en técnicas de imagen como la catoptroscopia, que mejoran la observación de estructuras cutáneas. Sin embargo, la disponibilidad limitada de especialistas y la heterogeneidad en las condiciones de captura entre distintos centros de salud han motivado el desarrollo de herramientas computacionales capaces de apoyar el análisis de imágenes dermatologías.

En los últimos años, el aprendizaje profundo ha impulsado avances significativos en segmentación de imágenes medicas, tarea que busca delimitar automáticamente la región de interés, por ejemplo la lesión, para facilitar etapas posteriores como medición, seguimiento y apoyo al diagnostico. Distintas lineas de investigación han propuesto arquitecturas de segmentación semántica y de instancias, con evaluaciones sobre conjuntos de datos estandarizados y desafíos competitivos. No obstante, gran parte de la evidencia reportada se construye sobre imágenes obtenidas en condiciones relativamente controladas o con distribuciones de datos homogéneas, lo que no siempre representa el escenario de uso real.

La relevancia de estudiar este problema radica en que, fuera del laboratorio, la cadena de adquisición introduce variaciones que afectan directamente la información disponible en la imagen. Cambios de iluminación, presencia de ruido, desenfoque, compresión y reducción de resolución pueden degradar la calidad de la señal de entrada

y, en consecuencia, alterar la salida del modelo. En contextos de recursos limitados, donde pueden utilizarse dispositivos de captura de menor costo o condiciones menos estandarizadas, esta variabilidad puede ser aun mas marcada. Por ello, mas alla de medir desempeño promedio en datos ideales, resulta necesario caracterizar la robustez del sistema y establecer condiciones mínimas de calidad que permitan una operación confiable.

A partir de lo anterior, el problema que aborda este trabajo es la falta de caracterización operacional sobre hasta que punto modelos de segmentación de lesiones cutáneas mantienen un comportamiento confiable cuando la imagen se degrada. El objetivo general es evaluar y comparar modelos de segmentación basados en aprendizaje profundo bajo degradaciones controladas de calidad de imagen, con el fin de identificar limites de operación y condiciones mínimas asociadas principalmente a resolución, ruido e iluminación. La naturaleza del estudio es experimental y comparativa, orientada a medir sensibilidad del desempeño ante perturbaciones sistemáticas en la entrada.

Para cumplir este objetivo, se implementan tres arquitecturas representativas (U-Net, DeepLabV3+ y Mask R-CNN) y se construye un entorno de evaluación que permite generar variantes degradadas de las imágenes mediante un esquema modular de transformaciones. La evaluación considera métricas de segmentación ampliamente utilizadas, tales como Intersección sobre Unión (IoU), coeficiente Dice, precisión y sensibilidad (recall), aplicadas sobre un conjunto de datos principal y una evaluación complementaria de generalización en un conjunto adicional. De este modo, se busca obtener evidencia cuantitativa sobre robustez y definir criterios técnicos que orienten el uso de estos modelos cuando la calidad de imagen no es ideal.

El documento se organiza de la siguiente manera. En el primer capitulo se presenta el contexto del problema y los objetivos del trabajo. Luego se revisan antecedentes y fundamentos relacionados con segmentación de lesiones cutáneas y arquitecturas de aprendizaje profundo. Posteriormente se describe la metodología, incluyendo los modelos evaluados, el banco de pruebas y el esquema de degradación. Finalmente se presentan y discuten los experimentos realizados, seguidos de las conclusiones y lineas de trabajo futuro.

1.1. Trabajos previos

La segmentación de imágenes médicas se ha consolidado como el componente tecnológico más crítico en los sistemas de asistencia al diagnóstico por computadora (CAD). Su propósito fundamental es la partición del dominio de la imagen en regiones que representen estructuras anatómicas o patológicas con coherencia semántica. A lo largo de la última década, la literatura científica ha documentado una transición profunda desde métodos de visión computacional clásica hacia arquitecturas de aprendizaje profundo de alta complejidad.

1.1.1. Segmentación en oncología y radiología de imágenes seccionales

La delimitación de tumores en imágenes de Resonancia Magnética (MRI) y Tomografía Computarizada (CT) representa uno de los mayores hitos de la disciplina. El desafío en este dominio radica en la alta variabilidad morfológica de los tejidos neoplásicos y el bajo contraste entre el núcleo tumoral y el edema circundante.

- **Neuro-oncología y el reto BraTS:** Los trabajos documentados por **Bakas et al. (2018)** [22] en el contexto del *Brain Tumor Segmentation Challenge* han servido como plataforma para la evolución de las arquitecturas 3D. Investigaciones previas han demostrado que el uso de *Convolutional Neural Networks* (CNN) con kernels tridimensionales permite capturar la continuidad espacial de los gliomas, superando las limitaciones de los modelos 2D tradicionales que procesaban cortes de forma aislada. Se ha observado que la integración de mecanismos de atención (*Attention Gates*) permite a los modelos priorizar regiones con gradientes de intensidad sutiles, mejorando la precisión en la identificación de los bordes del tumor.
- **Oncología Pulmonar:** En la segmentación de nódulos pulmonares en CT, trabajos como los de **Skourt et al. (2018)** [32] han validado la eficacia de la arquitectura U-Net. En este dominio, la literatura destaca la importancia del pre-procesamiento para normalizar las unidades Hounsfield, permitiendo que la red aprenda características de textura interna del nódulo que son imperceptibles

para el ojo humano, facilitando la distinción entre lesiones benignas y malignas en etapas tempranas.

1.1.2. Segmentación vascular en oftalmología

El análisis de la microvasculatura retiniana es esencial para el seguimiento de patologías como la retinopatía diabética y el glaucoma. Este campo presenta un desafío técnico particular: la segmentación de estructuras lineales delgadas que pueden tener apenas un píxel de ancho.

- **Modelos Recurrentes y Residuales:** El trabajo de **Alom et al. (2018)** [21] introdujo la arquitectura R2U-Net, la cual combina redes residuales con unidades recurrentes. Esta innovación permitió que los modelos recordaranç características de capas previas para mantener la conectividad de los vasos sanguíneos en los datasets DRIVE y STARE. La literatura en este campo enfatiza que la segmentación exitosa de la retina depende de la capacidad del modelo para gestionar el ruido de fondo y los reflejos de la cámara de fondo de ojo, utilizando capas de normalización por lotes (*Batch Normalization*) para estabilizar el aprendizaje.

1.1.3. Segmentación en gastroenterología y endoscopía

La detección de pólipos en colonoscopias es una tarea de segmentación en tiempo real que ha ganado tracción debido a su impacto en la prevención del cáncer colorrectal.

- **Sistemas de Atención Inversa:** El trabajo de **Fan et al. (2020)** [23] propuso la red **PraNet**, la cual utiliza un módulo de atención inversa para establecer una relación entre las características de bajo y alto nivel. La literatura técnica en este ámbito subraya que, a diferencia de la radiología, la endoscopía se enfrenta a un entorno dinámico con reflejos especulares y cambios constantes en la distancia focal. Los trabajos previos han recurrido a arquitecturas como HarDNet-MSEG para optimizar la velocidad de inferencia sin comprometer la precisión de la máscara, permitiendo la asistencia durante el procedimiento quirúrgico.

1.1.4. Segmentación en histopatología y nivel celular

La segmentación a nivel microscópico es el origen mismo de arquitecturas icónicas como la U-Net. En la patología digital, el objetivo es la delimitación de núcleos celulares y glándulas en placas teñidas con Hematoxilina y Eosina (H&E).

- **U-Net y UNet++:** Desde el trabajo seminal de **Ronneberger et al. (2015)** [20], la segmentación celular ha evolucionado hacia arquitecturas con rutas de agregación más densas. **Zhou et al. (2018)** [29] propusieron la UNet++, la cual introduce bloques densos en las conexiones de salto para reducir la brecha semántica entre el codificador y el decodificador. Estos trabajos han demostrado que la segmentación precisa de núcleos en tejidos densos es fundamental para el cálculo del índice mitótico, un marcador clave en la gradación del cáncer de mama.

1.1.5. Segmentación dermatológica y análisis de superficie

Finalmente, en el área de la dermatología, la segmentación de lesiones melanocíticas ha sido impulsada por el consorcio **ISIC**.

- **Evolución desde el Procesamiento Clásico:** Los trabajos previos de la década de los 2000 se basaban en métodos de *Level-Sets* y contornos activos (*snakes*). Sin embargo, la literatura moderna, liderada por autores como **Codella et al. (2018)**, ha demostrado que estas técnicas fallan ante la presencia de vello, burbujas de aire y variaciones cromáticas.
- **Arquitecturas Multiescala:** Investigaciones de **Bi et al. (2019)** han propuesto el uso de arquitecturas basadas en ResNet y estrategias de aprendizaje por etapas (*step-wise learning*). Estos trabajos destacan que la segmentación dermatológica requiere un equilibrio entre el contexto global (la piel circundante) y las características locales de la red de pigmento, validando el uso de módulos como el *Atrous Spatial Pyramid Pooling* (ASPP) para capturar información multiescala de manera eficiente.

1.1.6. Selección de modelos: Representatividad y robustez en el diagnóstico cutáneo

Tras analizar la evolución de la segmentación en diversos dominios médicos, se hace evidente que el éxito del diagnóstico asistido depende de la arquitectura de la red. Para esta investigación, se han seleccionado la U-Net, DeepLabv3+ y Mask R-CNN no solo por su alto desempeño bibliográfico, sino porque representan tres estrategias de ingeniería distintas —conexiones de salto, convoluciones dilatadas y detección basada en instancias— cuya robustez ante la degradación de imagen aún no ha sido comparada sistemáticamente.

1.1.6.1. U-Net: La red de simetría médica

Presentada por **Ronneberger et al.** [20], la U-Net es la arquitectura más utilizada en medicina debido a su capacidad de trabajar con pocos datos de entrenamiento.

Funcionamiento técnico: Posee una estructura simétrica de codificador (contracción) y decodificador (expansión). Su innovación principal son las *skip connections* o conexiones de salto, que transfieren información de alta resolución directamente a las capas de salida.

Explicación conceptual: Para un lector no técnico, la U-Net funciona como una letra Ü". En la bajada, la red resume"la imagen para entender qué está viendo (colores, texturas). En la subida, vuelve a agrandar la imagen para reconstruir la forma. Gracias a sus "puentes"de memoria (conexiones de salto), la red puede consultar sus notas originales para recordar exactamente dónde estaban los bordes finos que se pudieron borrar durante la compresión.

1.1.6.2. DeepLabv3+: Visión multiescala con zoom inteligente

Desarrollada por Google [1], esta red es reconocida por su eficiencia en el manejo de objetos de distintos tamaños.

Funcionamiento técnico: Utiliza convoluciones *atrous* (dilatadas) y el módulo *Atrous Spatial Pyramid Pooling* (ASPP). Esto permite aumentar el campo receptivo de los filtros sin perder resolución espacial.

Explicación conceptual: Imaginar mirar a través de un colador. En lugar de mirar píxeles pegados, esta red puede mirar puntos separados entre sí sin moverse de su posición. Esto le da un "zoom inteligente": puede entender simultáneamente el detalle de la lesión y el contexto de la piel sana que la rodea, sin necesidad de achicar la imagen original, lo que evita la pérdida de información crítica.

1.1.6.3. Mask R-CNN: Segmentación de instancias por etapas

Propuesta por el equipo de FAIR [2], esta arquitectura es la referencia para detectar y separar múltiples objetos individuales en una misma imagen.

Funcionamiento técnico: Trabaja sobre un paradigma de detección primero (*Region Proposal Network*) y segmentación después. Genera una máscara binaria para cada objeto detectado de forma independiente.

Explicación conceptual: Esta red actúa de forma muy humana en dos pasos. Primero, escanea la foto buscando "zonas de interész dibuja un cuadro (caja) alrededor de cada posible lesión. Segundo, toma ese cuadro y "pintaçon precisión de pincel la forma exacta de la mancha. Al separar el problema en "dónde estáz "qué forma tiene", resulta ser una arquitectura muy robusta y organizada para manejar imágenes complejas.

1.1.7. Discusión

La revisión de la literatura evidencia un panorama de contrastes en la segmentación de imágenes médicas. Por un lado, el campo ha alcanzado niveles de precisión sin precedentes en entornos controlados, pero por otro, enfrenta una "brecha de robustezçrítica al intentar trasladar estos algoritmos a la práctica clínica cotidiana [24, 8].

Éxito multidominio y elección de arquitecturas: La selección de las arquitecturas para este estudio no es arbitraria; responde a su éxito demostrado en diversos desafíos diagnósticos. La U-Net se ha consolidado como el estándar de oro en bioimagen debido a su diseño simétrico y sus çonexiones de salto"(*skip connections*), que permiten recuperar detalles espaciales finos que se pierden durante la compresión de datos, una característica vital tanto en la segmentación de células como en la delimitación de bordes de melanomas [20, 32].

Por su parte, DeepLabv3+ ha redefinido la segmentación semántica al introducir las convoluciones *atrous* (dilatadas) y el módulo ASPP, permitiendo que la red entienda el contexto global de una lesión sin sacrificar la resolución, una técnica que ha mostrado resultados superiores en la detección de pólipos y estructuras vasculares retinianas [1, 21, 23]. Finalmente, Mask R-CNN representa el estado del arte en la segmentación de instancias, permitiendo separar múltiples lesiones u objetos de forma independiente, una capacidad crucial validada en retos de neuroimagen como BraTS y en la detección de cáncer de piel en presencia de múltiples nevus [2, 22, 9].

El problema de la robustez en escenarios reales: A pesar de estas virtudes arquitectónicas, la literatura científica revela un problema recurrente: la mayoría de estos modelos han sido validados exclusivamente bajo condiciones clínicas ideales o en *datasets* curados como ISIC, donde las imágenes poseen estándares óptimos de iluminación, resolución y contraste [3, 10]. Existe un vacío significativo respecto al desempeño de estos sistemas ante la variabilidad del hardware de adquisición, especialmente en escenarios de baja infraestructura o telemedicina [19, 15].

Investigaciones en robustez de redes neuronales sugieren que los modelos entrenados en condiciones perfectas suelen sobreajustarse a texturas de alta frecuencia que el ojo humano ignora [25, 31]. Cuando estas redes se enfrentan a ruido de sensor, baja resolución o artefactos de post-procesamiento (típicos de cámaras de dispositivos móviles económicos), su capacidad de segmentación tiende a degradarse de forma no lineal.

Justificación del enfoque en telecomunicaciones: Desde la perspectiva de la telecomunicaciones, este problema se traduce en una degradación de la señal de entrada (la imagen) que corrompe la información semántica necesaria para el algoritmo. Los

trabajos previos han logrado avances significativos en la segmentación ïn-silico"(en el laboratorio), pero carecen de una cuantificación sistemática de los límites operacionales ïn-situ"(en el campo).

Justificamos, por tanto, el uso de U-Net, DeepLabv3+ y Mask R-CNN en esta investigación para evaluar cómo sus diferentes filosofías de diseño (conexiones de salto, convoluciones dilatadas y detección por etapas) responden ante la pérdida controlada de calidad de imagen. El objetivo es determinar si la sofisticación de estos modelos se traduce en una resiliencia real o si, por el contrario, su complejidad los hace más vulnerables ante las imperfecciones tecnológicas de los entornos rurales y la telemedicina de bajo costo [6, 11].

1.2. Definición del problema

La segmentación automática de lesiones cutáneas mediante técnicas de aprendizaje profundo (*Deep Learning*) constituye un desafío técnico de alta relevancia, dada la heterogeneidad morfológica de las lesiones y la variabilidad en las condiciones de adquisición [8, 10]. Si bien arquitecturas modernas como U-Net, DeepLabv3+ y Mask R-CNN han mostrado resultados prometedores en condiciones ideales de laboratorio, su desempeño se ve considerablemente afectado cuando las imágenes presentan degradaciones como baja resolución o ruido [20, 1, 2, 8].

1.2.1. Justificación técnica y económica

El problema central radica en la brecha existente entre los entornos de entrenamiento controlados y la realidad de la práctica clínica cotidiana [24, 8]. Gran parte de la investigación actual se centra en escenarios de captura ideales, sin considerar el impacto de la degradación de imagen en contextos reales [8, 11].

- **Disparidad de hardware:** Los modelos suelen ser entrenados con imágenes de alta calidad provenientes de equipos dermatoscópicos profesionales presentes en archivos como el de la ISIC [6, 3]. No obstante, en centros de atención con recursos limitados o en aplicaciones de telemedicina, el acceso a este equipamiento es restringido debido a sus altos costos [15, 19].

- **Variabilidad en cámaras reales:** La implementación práctica impone el uso de dispositivos de captura que varían drásticamente en calidad y precio, desde cámaras móviles de gama baja hasta sensores dermatoscópicos económicos [11, 15]. Actualmente, se desconoce qué tan simple o barata puede ser una cámara sin comprometer la eficacia del diagnóstico asistido [8, 19].
- **Degradación de la señal:** Desde una perspectiva de telecomunicaciones e ingeniería, el uso de hardware de bajo costo introduce ruido electrónico y térmico, además de una reducción en la resolución espacial [18, 25]. Estas imperfecciones actúan como interferencias que corrompen la señal visual, lo que puede provocar fallos críticos en la detección de la lesión al alejarse de la distribución de datos original [31, 25].
- **Artefactos de post-procesamiento:** El software interno de las cámaras comerciales a menudo aplica filtros de realce (*sharpening*) o suavizado que pueden ocultar características patológicas esenciales para la segmentación precisa o introducir texturas artificiales que confunden a la red [27, 14].

1.2.2. Finalidad y aplicación

La cuantificación sistemática de estos problemas tiene como propósito establecer criterios técnicos que aseguren la viabilidad de los modelos en entornos no ideales [11, 15].

- **Determinación de umbrales críticos:** Se busca identificar los niveles mínimos de resolución y los niveles máximos de ruido bajo los cuales la segmentación mantiene un desempeño confiable [8, 25]. Esto permite conocer si una captura de baja calidad permite aún una operación segura para el sistema [31].
- **Seguridad clínica:** Al definir los límites operativos, se reduce el riesgo de generar falsos negativos inaceptables causados por una señal de entrada deficiente, lo cual es crítico para la detección temprana de enfermedades como el melanoma [8, 9].
- **Democratización del diagnóstico:** Al establecer requerimientos mínimos de calidad para los dispositivos de captura, se facilita la adopción de herramientas precisas en entornos de recursos limitados, permitiendo diagnósticos

computacionales más inclusivos y resilientes frente a variaciones en los datos [19, 15].

1.3. Objetivos

1.3.1. Objetivo general

Determinar los umbrales de robustez y límites operativos críticos de las arquitecturas de segmentación de imágenes médicas U-Net, DeepLabv3+ y Mask R-CNN mediante el análisis sistemático de su desempeño ante la degradación controlada de imágenes dermoscópicas, con el fin de establecer los requerimientos técnicos mínimos de calidad de imagen que garanticen la viabilidad y seguridad del diagnóstico asistido por computadora en entornos de telemedicina basados en hardware de bajo costo.

Explicación y justificación del alcance: El desarrollo reciente de modelos de segmentación basados en deep learning ha logrado desempeños muy altos en bases de datos estandarizadas y en condiciones controladas. Sin embargo, al trasladar estos sistemas a escenarios clínicos reales, su rendimiento puede degradarse de forma significativa debido a variaciones en la adquisición de la imagen, tales como cambios de iluminación, diferencias de dispositivos, compresión, desenfoque, ruido y resoluciones heterogéneas. Esta brecha entre el desempeño reportado en laboratorio y la confiabilidad operativa en terreno constituye el problema central que aborda este proyecto.

Al finalizar este trabajo, se obtendrán los siguientes resultados:

- **Marco técnico de referencia para robustez:** Se determinarán umbrales de degradación a través de procesos como reducción de resolución, presencia de ruido y alteraciones de enfoque/iluminación en donde los modelos tendra una caida relevante en su desempeño. Esto permitirá pasar de una evaluación basada solo en promedios globales a criterios operativos verificables, útiles para definir límites de uso del sistema.
- **Evidencia para uso con imágenes de menor calidad:** Se hara un análisis orientado a escenarios de captura con recursos limitados, incluyendo la discusión

de qué modelo responde a ser mas viable para mantener un desempeño aceptable cuando las imágenes provienen de dispositivos de bajo costo o presentan condiciones de adquisición menos controladas.

En síntesis, el proyecto busca aportar criterios técnicos, basados en evaluación experimental, sobre la confiabilidad de modelos de segmentación ante degradaciones de calidad de imagen, como paso previo para considerar su uso en contextos médicos reales.

1.3.2. Objetivos específicos

- **Establecer el desempeño base de referencia:** Evaluar la capacidad de segmentación de las arquitecturas U-Net, DeepLabv3+ y Mask R-CNN utilizando imágenes en condiciones ideales (sin degradaciones) de los datasets ISIC 2018 y PH2. Este objetivo busca validar la selección de estas arquitecturas como *Gold Standard* y obtener métricas de referencia (Dice, Jaccard, Sensibilidad y Especificidad) contra las cuales contrastar el impacto de las corrupciones.
- **Analizar la resiliencia ante la pérdida de resolución espacial:** Determinar el impacto de la reducción sistemática de la resolución (desde el 100 % hasta el 3.125 %) en la precisión de la segmentación, identificando el punto de quiebre donde la pérdida de información geométrica compromete la detección del borde de la lesión.
- **Cuantificar la degradación por ruido aditivo de sensor:** Evaluar sistemáticamente la sensibilidad de los modelos ante diferentes niveles de ruido Gaussiano y Sal y Pimienta, analizando cómo la corrupción de la textura y el contraste afecta la estabilidad de los filtros convolucionales de cada red.
- **Correlacionar el diseño arquitectónico con la robustez operacional:** Comparar los resultados obtenidos para establecer cuál de las estrategias de ingeniería (conexiones de salto, convoluciones dilatadas o detección por etapas) ofrece una mayor tolerancia intrínseca ante las imperfecciones del hardware de captura de bajo costo.

1.4. Metodología

La metodología propuesta se basa en un enfoque experimental que simula condiciones adversas de captura para evaluar la **robustez** de los modelos de segmentación de lesiones cutáneas. El proceso se divide en cinco etapas principales, desde la preparación de los datos base hasta el análisis comparativo de los resultados.

1.4.1. Adquisición y preprocesamiento de datos

1. **Selección del *dataset* Base:** Se utilizará el *dataset* de la International Skin Imaging Collaboration (ISIC) 2018, el cual incluye imágenes dermatoscópicas y sus correspondientes máscaras de segmentación binarias.
2. **Uso de particiones oficiales:** Se utilizarán las particiones oficiales (**entrenamiento, validación y prueba**) tal como las proporciona el *dataset* ISIC 2018. Esto asegura que la evaluación de los modelos se realice bajo un estándar comparable y objetivo.
3. **Preprocesamiento estándar:** Todas las imágenes serán normalizadas y redimensionadas a un tamaño uniforme (según el requerimiento de cada arquitectura) para el entrenamiento base de los modelos.

1.4.2. Implementación y entrenamiento de arquitecturas

Se implementarán y entrenarán tres arquitecturas de segmentación, cada una con un enfoque diferente (segmentación semántica vs. de instancias):

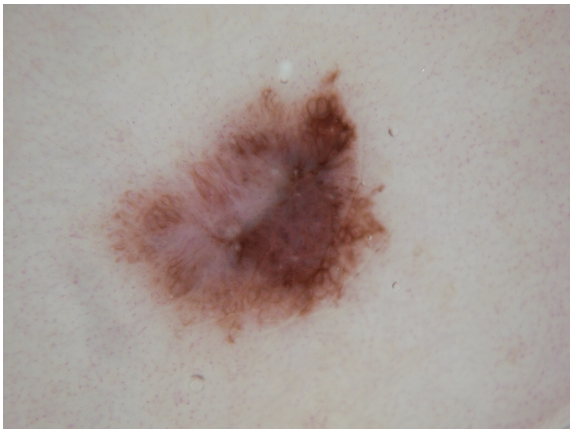
1. **DeepLabv3+ y U-Net (Segmentación semántica):** Implementadas con la librería ***segmentation_models*** sobre **TensorFlow/Keras**, utilizando *backbones* preentrenados (ej., ResNet34).
2. **Mask R-CNN (Segmentación de instancias):** Implementada con el *framework* **Detectron2** sobre **PyTorch** (ej., ResNet50-FPN).
3. **Entrenamiento base:** Los modelos serán entrenados y validados sobre el conjunto **sin degradación** para establecer sus métricas de desempeño (*baseline*)

y optimizar sus hiperparámetros.

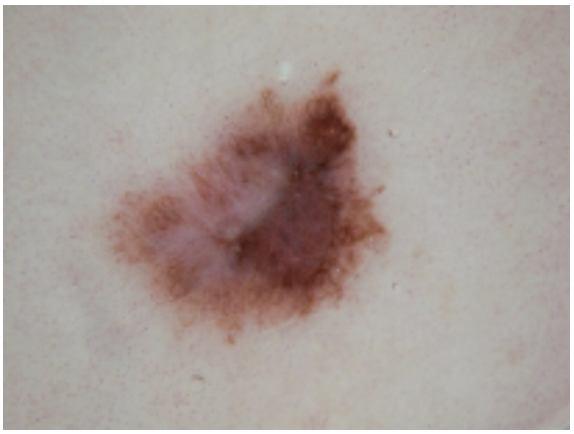
1.4.3. Diseño y aplicación del entorno de degradación sistemática

El proceso de degradación se aplica exclusivamente sobre el **conjunto de prueba (*test set*)** para generar *datasets* experimentales. Se definen cinco bloques generadores que utilizan las siguientes etiquetas de degradación:

1. **Bloque 1: Degradación de resolución exclusiva (`generate_block_1`):** Simula la pérdida de detalle espacial por *downsampling*. Se evalúan **10 niveles de resolución** indexados por porcentaje:
   - **`res_levels`:** $[3, 5, 8, 10, 15, 20, 25, 50, 80, 100]$ (% del tamaño original).



(a) Imagen dataset 2018 sin alteracion



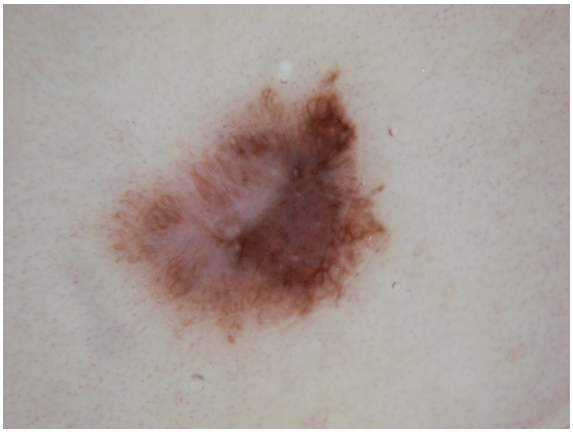
(b) Imagen dataset 2018 reducida a un 20 % con respecto a la resolucion original

Fig. 1.1: Comparación de bloque 1 (resolución).

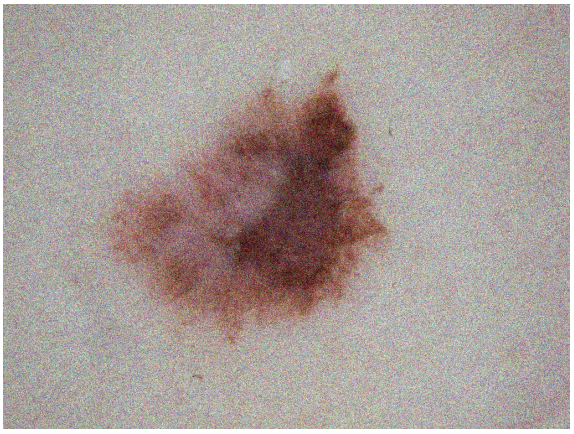
Los cambios de resolución representa la diferencia física entre un dermatoscopio profesional de alta resolución y una cámara web o teléfono antiguo. También simula la compresión agresiva necesaria para transmitir imágenes en entornos rurales con ancho de banda limitado. Al llegar al 3 %, buscamos el límite teórico donde la red deja de distinguir la forma macroscópica de la lesión.

2. **Bloque 2: Degradación de Ruido Exclusiva :** Aplica ruido sobre imágenes de resolución original.

- **noise_variants:** [ngaussian1, ngaussian3, ns&p1, ns&p3] (Ruido Gaussiano y Sal y Pimienta en niveles 0.01 y 0.03).



(a) Imagen dataset 2018 ISIC 0012591 sin alteracion



(b) Imagen dataset 2018 ISIC 0012591 con ruido gaussiano de varianza 0.03

Fig. 1.2: Comparación de bloque 2 (ruido).

El uso de ruido gaussiano simula el ruido térmico y electrónico generado por el sensor de imagen. Es común en dispositivos económicos con sensores pequeños que operan con alta sensibilidad ISO en condiciones de poca luz, mientras que Ruido Sal y Pimienta (s&p) representa errores puntuales en la digitalización o transmisión de datos. Se asocia a la pérdida de paquetes en redes móviles inestables o fallos físicos en píxeles individuales del sensor (dead pixels).

1.4.3.1. Justificación de los niveles de magnitud

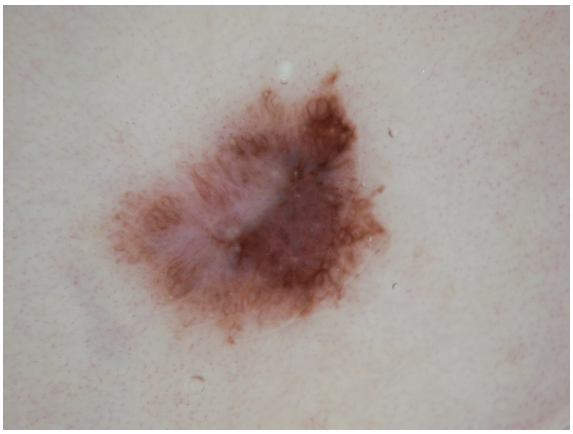
La selección de los valores de varianza (σ^2) para el ruido Gaussiano y de densidad de probabilidad para el ruido Sal y Pimienta en los niveles 0,01 y 0,03 no es arbitraria, sino que responde a la necesidad de representar dos estados críticos de la integridad de la señal de imagen en la cadena de telemedicina:

- El nivel 0.01 (*ngaussian1*) simula una captura con una ligera deficiencia de luz en un sensor de calidad estándar. El nivel 0.03 (*ngaussian3*) representa un escenario crítico de alto ruido térmico, permitiendo evaluar si la red puede generalizar y separar la patología de la estática electrónica extrema [25, 31]

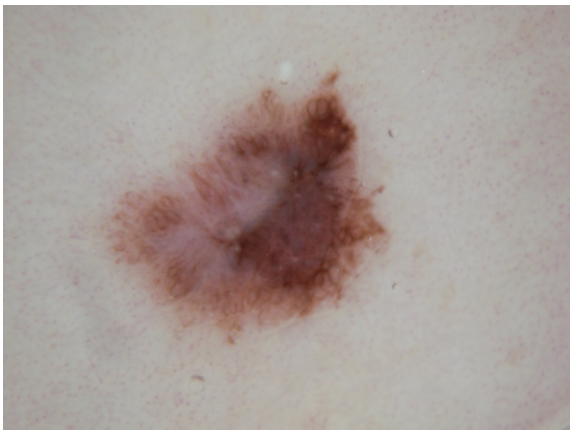
- Para el Ruido Sal y Pimienta los niveles 0.01 y 0.03 permiten medir la robustez ante la perforación de la señal visual, determinando si las redes son capaces de interpolar la forma de la lesión a pesar de la pérdida de información local de contraste.

3. **Bloque 3: Degradación de filtros exclusiva :** Aplica algoritmos de post-procesamiento sobre la imagen original.

   - **`filter_levels:`** [`blur`, `sharpen`, `blur_sharpen`] (Suavizado, realce y su combinación).



(a) Imagen dataset 2018 ISIC 0012591 sin alteración



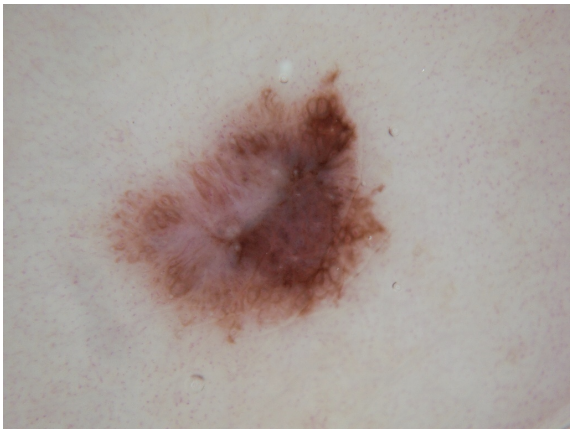
(b) Imagen dataset 2018 ISIC 0012591 con filtro blur aplicado

Fig. 1.3: Comparación de bloque 3 (filtro).

El filtro de Suavizado (Blur) simula el desenfoque por movimiento o el post-procesamiento para reducción de ruido que elimina detalles finos de la textura de la piel, mientras que el filtro de Realce (Sharpen) simula algoritmos comerciales que acentúan bordes. Se justifica su estudio porque estos filtros pueden amplificar el ruido de fondo o crear texturas artificiales que generan falsos positivos en los modelos de aprendizaje profundo.

4. **Bloque 4: Combinación de resolución y ruido:** Combina la pérdida de detalle espacial y el ruido.

   - **Resolución:** $[40, 60, 80]$.
   - **Ruido:** [`ngaussian1`, `ngaussian3`, `ns&p1`, `ns&p3`].



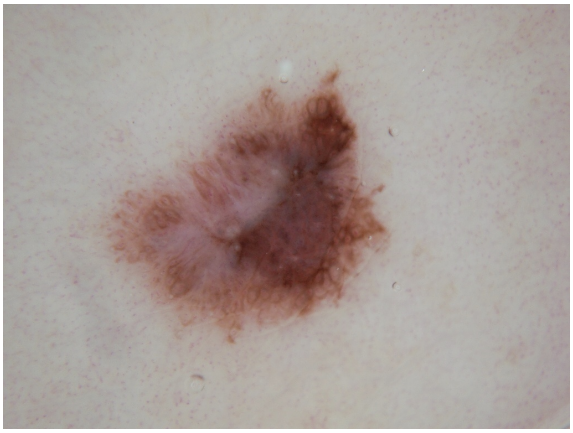
(a) Imagen dataset 2018 ISIC 0012591 sin alteracion

(b) Imagen dataset 2018 ISIC 0012591 reducida a una resolucion de 60 % un con ruido gaussiano de varianza 0.03

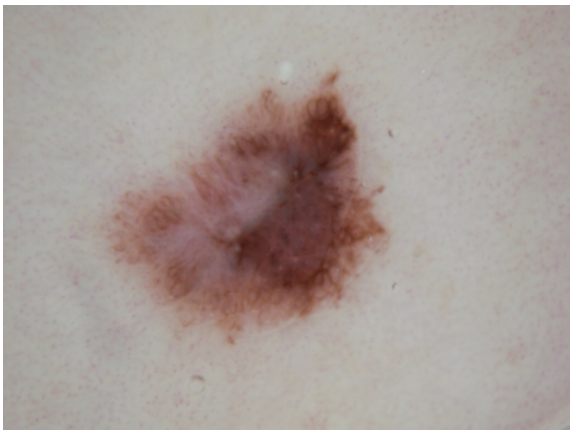
Fig. 1.4: Comparación de bloque 4 (resolución + ruido).

Permite aislar el efecto combinado de un hardware sensor deficiente (ruido) y una transmisión limitada por ancho de banda (baja resolución). El proposito de este bloque es determinar si los modelos pueden compensar la falta de resolución mediante la interpolación de características, o si el ruido satura los filtros convolucionales impidiendo la segmentación.

5. **Bloque 5: Combinación de resolución + filtro:** Combina la pérdida de detalle espacial sumado a la aplicación de filtros.
   - **Resolución:** $[40, 60, 80]$.
   - **Ruido:** [`ngaussian1`, `ngaussian3`, `ns&p1`, `ns&p3`].
   - **Filtros:** [`blur`, `sharpen`, `blur_sharpen`].



(a) Imagen dataset 2018 ISIC 0012591 sin alteracion

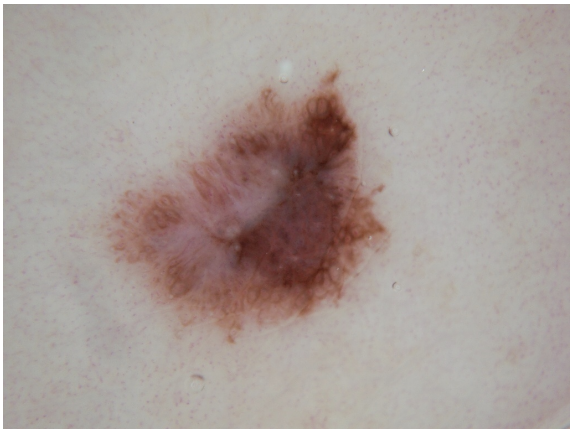


(b) Imagen dataset 2018 ISIC 0012591 reducida a una resoluccion de 60 % con filtro blur aplicado

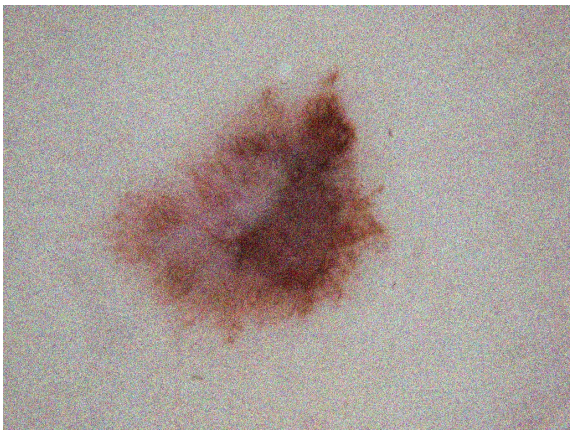
Fig. 1.5: Comparación de bloque 5 (resolución + filtro).

Representa la salida típica de dispositivos móviles de gama baja que, para compensar la falta de nitidez debida a un sensor pequeño o a una óptica deficiente, aplican filtros de *sharpening* (afilado) de forma agresiva, con el fin de evaluar si el realce artificial de bordes en una señal de baja resolución introduce artefactos de *aliasing* o "bordes falsos"que puedan inducir a la red a generar falsos positivos o a distorsionar la geometría real de la lesión.

6. **Bloque 6: Combinación de Ruido + Filtro:** Combina la aplicación de ruido y los filtros.
   - **Ruido:** [`ngaussian1`, `ngaussian3`, `ns&p1`, `ns&p3`].
   - **Filtros:** [`blur`, `sharpen`, `blur_sharpen`].



(a) Imagen dataset 2018 ISIC 0012591 sin alteracion

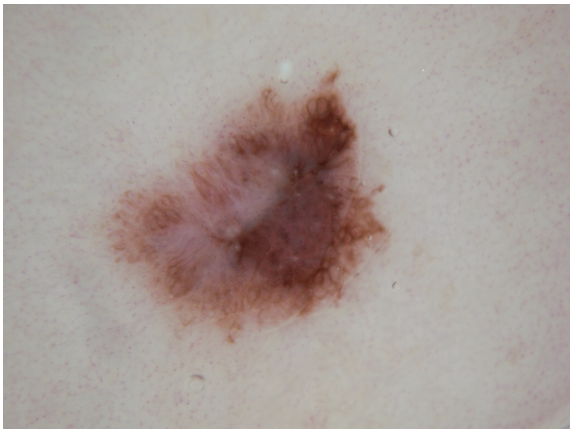


(b) Imagen dataset 2018 ISIC 0012591 con ruido gaussiano de varianza 0.03 con filtro blur sharpen aplicado

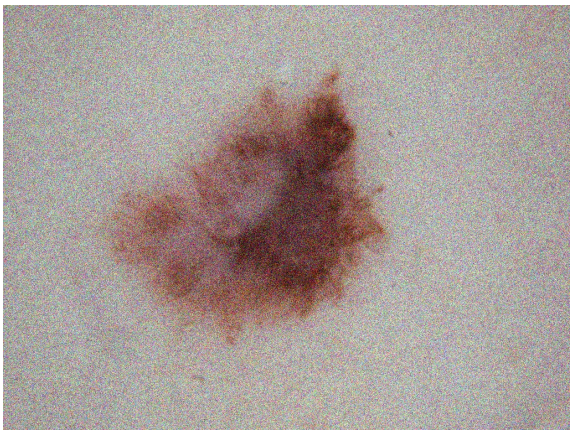
Fig. 1.6: Comparación de bloque 6 (ruido + filtro).

Físicamente, este escenario es extremadamente común en cámaras comerciales que operan en condiciones de baja luminosidad. El software de la cámara intenta mitigar el ruido del sensor mediante filtros de suavizado (*blur*) o, por el contrario, intenta recuperar el detalle perdido aplicando filtros de realce sobre una señal ya corrompida, esto es con el fin de cuantificar el impacto del ruido procesado". Desde la ingeniería, es crítico determinar si un filtro de suavizado ayuda a la red al eliminar el ruido Gaussiano, o si por el contrario, perjudica la segmentación al borrar las texturas sutiles de la lesión que son necesarias para un diagnóstico preciso.

7. **Bloque 7: Combinación triple (resolución + ruido + filtros):** Simula las condiciones adversas más realistas.
   - **Resolución:** $[40, 60, 80]$.
   - **Ruido:** [`ngaussian1`, `ngaussian3`, `ns&p1`, `ns&p3`].
   - **Filtros:** [`blur`, `sharpen`, `blur_sharpen`].



(a) Imagen dataset 2018 ISIC 0012591 sin alteracion



(b) Imagen dataset 2018 ISIC 0012591 reducido a una resolucion de 60 % con ruido gaussiano de varianza 0.03 con filtro blur sharpen aplicado

Fig. 1.7: Comparación de bloque 7 (resolución + ruido + filtros).

Representa el escenario clínico más realista y hostil. En la práctica, una imagen capturada en una zona remota suele sufrir de baja resolución (por el dispositivo), ruido (por la mala iluminación) y distorsión por filtros automáticos (del software del teléfono), esto se hace con el fin de evaluar la sinergia de los artefactos y determinar si el sistema mantiene un nivel de seguridad clínica mínimo bajo estrés extremo.

1.4.4. Evaluación de la Robustez y Análisis Comparativo

1. **Evaluación sobre *Test Sets* Degradados:** Los modelos serán evaluados sobre **todos** los *datasets* generados en la Sección 1.4.3 para medir la caída de desempeño.
2. **Métricas clave:** Se registrarán el Dice Coefficient, la Intersección sobre Unión (IoU), la sensibilidad (recall) y la precision para cada escenario de evaluación.
3. **Análisis de umbrales críticos:** Se comparará el desempeño de U-Net, DeepLabv3+ y Mask R-CNN frente a cada degradación para identificar el límite operativo de cada arquitectura y determinar cuál presenta mayor robustez.

1.5. Alcances y limitaciones

1.5.1. Alcances

El alcance del presente proyecto se centra en la evaluación comparativa y sistemática de la robustez de arquitecturas de *Deep Learning* para la segmentación de lesiones cutáneas. Los principales logros que se esperan obtener incluyen:

1. **Evaluación comparativa de arquitecturas clave:** Se implementarán y compararán tres modelos de alto desempeño y amplia adopción en el ámbito médico: **U-Net**, **DeepLabv3+** (segmentación semántica) y **Mask R-CNN** (segmentación de instancias).
2. **Determinación de umbrales operacionales:** Se determinarán experimentalmente los límites de resistencia a la degradación de cada modelo, identificando los umbrales críticos para **10 niveles de resolución** (hasta 3 % del tamaño original), **4 variantes de ruido** (Gaussiano y Sal y Pimienta) y **3 tipos de filtros** de post-procesamiento.
3. **Análisis de robustez combinada:** Se realizará una evaluación rigurosa en escenarios de degradación combinada para simular condiciones clínicas realistas, aportando datos valiosos sobre la sinergia de los artefactos de imagen.
4. **Generación de lineamientos técnicos:** Los resultados permitirán emitir recomendaciones precisas sobre los requerimientos mínimos de calidad para los dispositivos de captura utilizados en sistemas de diagnóstico dermatológico automatizado.
5. **Uso de Entornos Validados:** El estudio se realizará sobre las particiones oficiales (*training, validation* y *test*) del *dataset* ISIC 2018, asegurando la imparcialidad de la evaluación de robustez.

1.5.2. Limitaciones

Las limitaciones inherentes al diseño y el alcance acotado de una tesis de pregrado incluyen:

1. **Limitación de arquitecturas:** El estudio se centrará exclusivamente en las tres arquitecturas seleccionadas (U-Net, DeepLabv3+, Mask R-CNN), **no abarcando otras variantes** de modelos de segmentación o arquitecturas emergentes.
2. **Uso de *datasets* públicos:** Aunque el *dataset* ISIC 2018 es un estándar, la variabilidad en un contexto clínico real de rutina o la diversidad de tipos de piel y lesiones podrían ser mayores a las contenidas en el conjunto de datos de prueba público.
3. **Limitación en factores clínicos:** El análisis se limitará a condiciones simuladas de degradación de imagen (resolución, ruido, filtros, iluminación) y no considerará otros factores clínicos o logísticos como tipos de piel, artefactos de captura no simulados, o la variabilidad inter-observador en la generación de las máscaras de verdad fundamental.
4. **Generalización de resultados:** Los resultados de robustez serán estrictamente válidos para las arquitecturas y los *backbones* específicos utilizados en la implementación (ej., ResNet34/ResNet50-FPN), y su extrapolación a otras configuraciones requerirá una validación adicional.

Capítulo 2

Marco Teórico

2.1. Introducción a la dermatoscopía y el diagnóstico asistido

La detección temprana de cáncer de piel, particularmente el melanoma, es crucial para la supervivencia del paciente. Históricamente, el diagnóstico se ha basado en la inspección visual (regla ABCDE) y la **dermatoscopía**, una técnica no invasiva que utiliza un microscopio de superficie y luz polarizada para examinar estructuras subepidérmicas de la lesión. La interpretación de estas imágenes dermatoscópicas requiere de un alto grado de experiencia por parte del dermatólogo.

La necesidad de automatizar y estandarizar este proceso ha impulsado el uso de sistemas de diagnóstico asistido por computadora (CAD, por sus siglas en inglés), siendo el *Deep Learning* la tecnología de vanguardia para este propósito.

2.2. Fundamentos del aprendizaje profundo (*Deep Learning*)

El *Deep Learning* es una subrama del Aprendizaje Automático (Machine Learning) basada en redes neuronales artificiales con múltiples capas (*deep neural networks*).

Estas redes son capaces de aprender representaciones jerárquicas de los datos de forma automática, sin necesidad de extracción manual de características.

2.2.1. Redes neuronales convolucionales (CNNs)

Las Redes Neuronales Convolucionales (CNNs) constituyen la arquitectura fundamental para el procesamiento de señales visuales en el dominio del *Deep Learning*. Su eficacia se basa en la capacidad de aprender una **jerarquía de características** de forma automática: las capas iniciales detectan estructuras simples (bordes, gradientes de color), mientras que las capas profundas integran esta información para reconocer patrones semánticos complejos (formas de lesiones, texturas patológicas). El éxito de esta jerarquía depende de tres componentes críticos:

1. **Capa convolucional (*Convolutional layer*):** Aplica un conjunto de filtros (*kernels*) a la imagen de entrada para producir mapas de características (*feature maps*). Esto permite que la red aprenda patrones espaciales y texturas, como bordes, colores y formas, cruciales en la dermatoscopía.
2. **Capa de activación (*Activation layer*):** Generalmente se usa la función **ReLU** (Rectified Linear Unit) para introducir no linealidad en el sistema, lo que permite que la red aprenda funciones complejas.
3. **Capa de *pooling*:** Reduce la dimensionalidad espacial de los mapas de características, disminuyendo la cantidad de parámetros y haciendo la red más robusta a pequeñas traslaciones de los objetos en la imagen.

2.3. Segmentación de imágenes médicas

La **segmentación de imágenes** es una tarea de visión por computadora que consiste en dividir una imagen digital en múltiples segmentos, asignando una etiqueta de clase a cada *píxel*. En el contexto de lesiones cutáneas, el objetivo es aislar la lesión (primer plano) del tejido sano circundante (fondo).

2.3.1. Segmentación semántica vs. Segmentación de instancias

1. **Segmentación Semántica:** Asigna una etiqueta de clase a cada píxel, sin distinguir entre instancias individuales del mismo objeto. En el caso de una única lesión, esto se traduce en clasificar cada píxel como "lesión.º "fondo". Las arquitecturas como **U-Net** y **DeepLabv3+** se centran en esta tarea.
2. **Segmentación de Instancias:** Identifica cada objeto individualmente y asigna una máscara de píxeles a cada instancia. Si hubiera múltiples lesiones en una sola imagen, segmentación de instancias las distinguiría por separado. **Mask R-CNN** es un ejemplo de esta aproximación.

2.4. Arquitecturas clave para la segmentación

2.4.1. U-Net: Arquitectura de codificador-decodificador

Propuesta por Ronneberger *et al.* en 2015 [20], U-Net es la arquitectura canónica para la segmentación de imágenes biomédicas. Su diseño se caracteriza por:

1. **Trayectoria de contracción (codificador):** Sigue la estructura típica de una CNN para capturar el contexto de la imagen.
2. **Trayectoria expansiva (decodificador):** Utiliza capas de convolución transpuesta (o *upsampling*) para aumentar progresivamente la resolución de los mapas de características.
3. **Conexiones de salto (*skip connections*):** Estas conexiones directas unen las características del codificador con el decodificador en el mismo nivel de resolución. Esto es crucial para transferir información de alto detalle espacial (bordes finos) que se pierde en la contracción, permitiendo una localización de píxeles muy precisa.

2.4.2. DeepLabv3+: Segmentación con convoluciones atrous

Desarrollada por Google, DeepLabv3+ es una de las arquitecturas líderes en segmentación semántica. Sus innovaciones principales son:

1. **Convolución *atrous* (dilatada):** Permite expandir el campo de visión de los filtros sin aumentar el número de parámetros ni reducir la resolución espacial. Es decir, captura información de contexto más amplia sin perder el detalle del objeto.
2. **Spatial Pyramid Pooling (ASPP):** Módulo que aplica convoluciones *atrous* con diferentes tasas de dilatación. Esto captura información de objetos en múltiples escalas, mejorando el rendimiento en lesiones con gran variabilidad de tamaño.
3. **Estructura de codificador-decodificador mejorada:** Combina las características ricas en semántica del codificador con información de bordes y localización del decodificador.

2.4.3. Mask R-CNN: Segmentación basada en detección (instancias)

Mask R-CNN es una extensión del algoritmo Faster R-CNN, que además de detectar y clasificar objetos, genera una máscara de segmentación precisa para cada instancia detectada. Funciona en tres etapas:

1. **Red de propuesta de región (RPN):** Identifica las regiones candidatas donde podría haber un objeto.
2. **Clasificación y *bounding box*:** Para cada región candidata, se predice la clase del objeto y su caja delimitadora (detección).
3. **Generación de máscara:** Paralelamente a la clasificación, se añade una rama para predecir la máscara de segmentación a nivel de píxel para la instancia detectada.

2.5. Robustez en sistemas de visión por computadora

La robustez se define como la capacidad de un sistema para mantener su integridad operativa ante datos que presentan desviaciones estadísticas respecto al conjunto de entrenamiento. En telemedicina, esta capacidad es crítica debido a la falta de control sobre el hardware de adquisición.

2.5.1. Fenomenología de la degradación de la señal de imagen

Las degradaciones simuladas en este estudio representan fenómenos físicos reales en la cadena de captura y transmisión:

1. **Reducción de resolución (*downsampling*):** Representa la limitación de ancho de banda en la transmisión y la baja densidad de píxeles en sensores económicos. Provoca una pérdida de la señal de alta frecuencia, suavizando los bordes críticos para el diagnóstico.
2. **Ruido gaussiano (ruido térmico):** Simula la interferencia electrónica inherente a los sensores CMOS, especialmente en condiciones de baja luminosidad donde la relación señal-ruido (SNR) se degrada significativamente.
3. **Ruido sal y pimienta (ruido de transmisión):** Representa errores de cuantización o pérdida de paquetes durante la transmisión digital, manifestándose como impulsos de intensidad extrema que corrompen la coherencia local de la imagen.
4. **Filtros de post-procesamiento:** Simulan algoritmos comerciales de realce (*Sharpen*) o suavizado (*Blur*) que, aunque mejoran la estética visual, pueden distorsionar las características semánticas originales que la red neuronal necesita para segmentar con precisión.

2.5.2. Degradación de la imagen: ruido, resolución y filtros

Las degradaciones de la imagen son el foco experimental de este estudio, ya que simulan condiciones adversas en un entorno de telemedicina o de baja infraestructura.

1. **Reducción de resolución (*Downsampling*):** Simula la captura con cámaras de baja resolución. Una reducción extrema causa la pérdida irreversible de detalles finos y la ambigüedad en los bordes de la lesión.
2. **Ruido aditivo:** Se introduce como resultado de errores en el sensor o en la transmisión. Se distinguen dos tipos clave:
   - **Ruido gaussiano:** Variación estadística aleatoria que sigue una distribución normal.
   - **Ruido sal y Pimienta (*Salt and Pepper*, s&p):** Píxeles que toman valores extremos (blanco o negro) de forma aleatoria.
3. **Filtros de procesamiento:** Simulan el post-procesamiento realizado por el dispositivo.
   - **Suavizado (*Blur*):** Difumina los bordes y reduce el ruido, pero a costa de la pérdida de detalles finos.
   - **Realce (*Sharpen*):** Acentúa los bordes, pero puede introducir artefactos no deseados en la lesión.

2.6. Métricas de evaluación de segmentación

Para cuantificar el desempeño de las arquitecturas de segmentación y evaluar su robustez, es necesario comparar la máscara predicha por el modelo (P) con la máscara de verdad fundamental o *Ground Truth* (GT) proporcionada por los especialistas [8, 10]. Esta comparación se basa en la clasificación de píxeles en cuatro categorías: Verdaderos Positivos (TP), Verdaderos Negativos (TN), Falsos Positivos (FP) y Falsos Negativos (FN) [27].

2.6.1. Coeficiente Dice (*Dice Coefficient*)

El Coeficiente Dice, también conocido como puntaje F1, es la métrica de referencia en la segmentación de imágenes médicas debido a su capacidad para manejar el desbalance de clases, donde el área de la lesión suele ser mucho menor que el fondo sano [8, 27].

$$\text{Dice} = \frac{2 \times |P \cap GT|}{|P| + |GT|} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.1}$$

Funcionamiento y lógica: La fórmula mide la similitud entre dos conjuntos duplicando el peso de la intersección (los píxeles donde el modelo y el especialista coinciden) y dividiéndolo por la suma total de píxeles identificados como positivos por ambos [27]. Un valor de 1 indica una segmentación perfecta, mientras que 0 indica una ausencia total de solapamiento. Al ignorar los Verdaderos Negativos (TN), el Dice se enfoca exclusivamente en la precisión de la delimitación de la lesión.

2.6.2. Intersección sobre unión (IoU o Índice Jaccard)

El índice IoU es una métrica geométrica que mide el grado de solapamiento entre la máscara predicha y la real.

$$\text{IoU} = \frac{|P \cap GT|}{|P \cup GT|} = \frac{TP}{TP + FP + FN} \tag{2.2}$$

Funcionamiento y lógica: A diferencia del Dice, el IoU penaliza de forma más severa los errores de segmentación (píxeles clasificados incorrectamente) al dividir la intersección de las áreas por su unión total. Matemáticamente, siempre resulta en un valor ligeramente menor o igual al Dice, proporcionando una medida más conservadora de la exactitud espacial del modelo.

2.6.3. Precisión (*Precision*)

La precisión mide la fidelidad del modelo al identificar píxeles como parte de la lesión, indicando qué proporción de las predicciones positivas son realmente correctas.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.3}$$

Funcionamiento y lógica: La fórmula relaciona los aciertos (TP) con el total de píxeles que el modelo çree"que son lesión ($TP + FP$). Una precisión alta indica que el modelo es cauteloso y no tiende a sobre-segmentar o incluir tejido sano dentro de la máscara de la lesión. En entornos de telemedicina con hardware de bajo costo, esta métrica es vital para evitar el diagnóstico erróneo de áreas sanas como sospechosas.

2.6.4. Sensibilidad (*Recall* o *Sensitivity*)

La sensibilidad cuantifica la capacidad del modelo para detectar la totalidad de la región patológica presente en la imagen original.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.4}$$

Funcionamiento y lógica: Esta métrica divide los aciertos (TP) por el área real total de la lesión ($TP + FN$). Un valor alto de sensibilidad es el objetivo primordial en oncología cutánea, ya que indica que el modelo minimiza los Falsos Negativos (FN); es decir, asegura que ninguna parte de la lesión maligna pase inadvertida por el sistema de detección automática [8, 10].

Capítulo 3

Desarrollo

3.1. Introducción

El Capítulo 3 describe las actividades de implementación práctica y la configuración experimental llevadas a cabo para cumplir con los objetivos del proyecto. Esto incluye la selección y configuración del entorno de trabajo, la adaptación de las arquitecturas de *Deep Learning* y, centralmente, el desarrollo del sistema modular de degradación de imágenes.

3.2. Configuración del entorno de trabajo

3.2.1. Plataforma de hardware y software

La experimentación se ejecutó en una plataforma de alto rendimiento para mitigar los tiempos de entrenamiento, utilizando la siguiente configuración de *hardware* y *software* base:

- **GPU:** NVIDIA GeForce RTX 3070 con 8 GB de VRAM.
- **Sistemas operativos:** Entorno dual basado en Ubuntu para la ejecución de Jupyter/Python.
- **Frameworks de *Deep Learning*:**

- **TensorFlow 2.x/Keras:** Utilizado para la implementación de U-Net y DeepLabv3+.
- **PyTorch 2.x/CUDA 11.8:** Utilizado para la implementación de Mask R-CNN [16].

3.2.2. Dataset Utilizado

Se utilizó el International Skin Imaging Collaboration (ISIC) 2018 [3]. Este conjunto de datos proporciona imágenes dermatoscópicas y las correspondientes máscaras de segmentación binarias (verdad fundamental). Se respetaron las particiones oficiales de entrenamiento, validación y prueba (*test set*) proporcionadas, asegurando la objetividad en la evaluación de la robustez.

3.3. Implementación de arquitecturas de segmentación

Se implementaron tres modelos, cada uno con un enfoque de segmentación distinto:

1. **U-Net** [20]: Implementada con *segmentation_models* y *backbones* preentrenados (ej., ResNet34) en ImageNet.
2. **DeepLabv3+** [1]: Implementada con *segmentation_models* y configurada para aprovechar la convolución *atrous*.
3. **Mask R-CNN** [2]: Implementada a través de **Detectron2**, adaptada para procesar imágenes y máscaras en el formato de Segmentación de Instancias.

El entrenamiento base de estos modelos se realizó sobre las imágenes originales de los conjuntos de entrenamiento y validación sin degradación para establecer las métricas de desempeño iniciales (*baseline*).

3.4. Desarrollo del sistema modular de degradación

El componente central del proyecto es el módulo `Degradaciones_Variadas_generador.py`, diseñado para simular de manera sistemática y controlada las condiciones adversas de captura. Este sistema opera exclusivamente sobre el conjunto de prueba (*test set*) del ISIC para generar los *datasets* experimentales.

3.4.1. Bloques generadores y parámetros de prueba

El sistema se compone de cinco bloques generadores (implementados en `Degradaciones_Variadas_generador.py`) que cubren distintos tipos de artefactos:

1. **Bloque 1: Resolución exclusiva (`generate_block_1`):** Evalúa el impacto de la pérdida de detalle espacial por *downsampling*. Se prueban 10 niveles desde el 100 % hasta el 3 % del tamaño original $(3, 5, 8, 10, 15, 20, 25, 50, 80, 100)$. Este bloque busca determinar el límite teórico donde la pérdida de información geométrica impide la identificación de la lesión.

2. **Bloque 2: Ruido exclusivo (`generate_block_2`):** Aplica interferencias estadísticas sobre la resolución original (100 %). Incluye Ruido Gaussiano (`ngaussian`) y Ruido Sal y Pimienta (`ns&p`) en magnitudes de varianza y densidad de 0,01 (Nivel 1) y 0,03 (Nivel 3). El nivel 0,01 simula ruido base de sensores CMOS estándar, mientras que el 0,03 representa un escenario crítico de baja luminosidad y alta ganancia ISO [31, 25].

3. **Bloque 3: Filtros exclusivos (`generate_block_3`):** Evalúa el impacto de algoritmos de post-procesamiento como el suavizado (*blur*), el realce de bordes (*sharpen*) y su aplicación combinada sobre la imagen original. Estos filtros simulan las correcciones automáticas realizadas por el software de dispositivos móviles comerciales.

4. **Bloque 4: Combinación doble - resolución + ruido (`generate_block_4`):** Analiza la sinergia entre la pérdida de píxeles y el ruido del sensor. Cruza los niveles de resolución de operación realista $(40, 60, 80\,\%)$ con las cuatro variantes

de ruido. Se justifica el uso de estos porcentajes de resolución para representar la compresión estándar en redes de telemedicina con ancho de banda limitado.

5. **Bloque 5: Combinación doble - resolución + filtro (`generate_block_5`):** Evalúa cómo el procesamiento digital afecta a una señal que ya ha perdido resolución espacial. Busca detectar si el realce artificial de bordes (*sharpening*) en baja resolución introduce artefactos que confunden a los filtros convolucionales de las redes.
6. **Bloque 6: Combinación doble - ruido + filtro (`generate_block_6`):** Analiza si los filtros de suavizado mitigan el ruido electrónico o si, por el contrario, perjudican la segmentación al borrar texturas sutiles necesarias para el diagnóstico. Representa el compromiso típico entre la limpieza de la imagen y la preservación de detalles patológicos.
7. **Bloque 7: Combinación triple (`generate_block_7`):** Representa el escenario más hostil y realista de la investigación. Combina simultáneamente tres niveles de resolución $(40, 60, 80\,\%)$, las cuatro variantes de ruido y los tres tipos de filtros, resultando en 36 escenarios de degradación severa. Se utiliza para determinar la sinergia negativa entre artefactos en entornos de baja infraestructura.

3.5. Proceso de Evaluación Cuantitativa

3.5.1. Ejecución de las pruebas y cálculo de métricas

La evaluación se centralizó en el script `main.py`, donde se implementó la lógica de carga, predicción y evaluación.

1. **Carga del modelo:** Cada modelo fue cargado en su respectivo entorno.
2. **Evaluación Iterativa:** Se iteró sobre cada uno de los *datasets* degradados.
3. **Cálculo de métricas:** La función de evaluación calculó y almacenó para cada predicción las métricas de segmentación: **Dice**, **IoU**, **Sensibilidad (*Recall*)** y **Precisión (*Precision*)**.

4. **Almacenamiento de Resultados:** Los resultados fueron almacenados en archivos CSV (`resultados_..._.csv`) para el análisis detallado posterior.

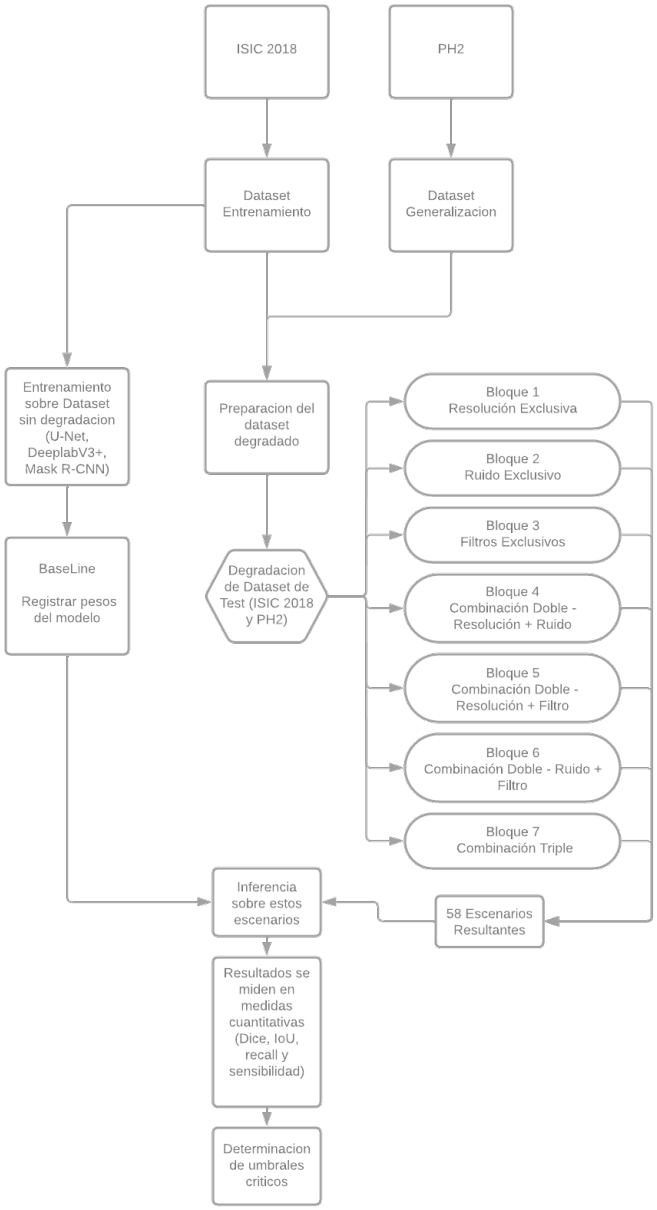


Fig. 3.1: Diagrama de flujo del proceso.

Capítulo 4

Resultados

4.1. Introducción

En este capítulo se reportan y analizan los resultados de desempeño y robustez operacional de tres arquitecturas de segmentación: U-Net, DeepLabv3+ y Mask R-CNN. La evaluación se estructura en tres niveles:

- **Desempeño base (sin degradación):** establece la referencia de calidad máxima posible sobre ISIC 2018 y permite comparar equilibrio entre métricas.
- **Robustez global (con degradación):** resume el efecto promedio de familias de degradación (resolución, ruido y filtros), y de combinaciones de estas, sobre el conjunto de prueba de ISIC 2018.
- **Umbrales operacionales:** identifica condiciones limite donde la segmentación deja de ser confiable. Para ello se analizan gráficos de dispersion (Intersection over Union, IoU, y coeficiente de Dice, Dice Similarity Coefficient, DSC) junto con métricas de precisión (Precision) y exhaustividad (Recall).

Para facilitar la lectura, se usan dos tipos de visualizaciones: (i) gráficos radar para comparar la "huella" global de cada modelo en IoU, Dice, Precision y Recall, y (ii) gráficos de dispersion y barras para ubicar degradaciones especificas y evidenciar trade-offs (por ejemplo, alta Recall con baja Precision sugiere sobre-segmentación).

4.2. Métricas de desempeño base y generalización

4.2.1. Desempeño base sobre ISIC 2018 (sin degradación)

Tabla 4.1: Comparación de métricas por modelo sobre ISIC 2018

Modelo	IoU promedio	Dice promedio	Precision promedio	Recall promedio
U-Net	0.8284	0.893	0.9156	0.9044
DeepLabv3+	0.7867	0.8631	0.8995	0.8733
Mask R-CNN	0.788	0.8688	0.8351	0.9445

A partir de la Tabla 4.1 se observan diferencias relevantes entre desempeño promedio y el tipo de error que comete cada arquitectura. U-Net obtiene el mejor desempeño base (IoU = 0.8284, Dice = 0.893), con un balance favorable entre Precision (0.9156) y Recall (0.9044), lo que sugiere segmentaciones consistentes sin tendencia marcada a sobre o sub-segmentar.

DeepLabv3+ presenta una disminución respecto a U-Net (IoU = 0.7867, Dice = 0.8631), manteniendo Precision alta (0.8995) pero con menor Recall (0.8733). En términos prácticos, este patrón suele asociarse a predicciones mas conservadoras (menos falsos positivos) a costa de perder parte del borde o zonas de baja contrastación.

Mask R-CNN alcanza un IoU similar a DeepLabv3+ (0.788) y Dice (0.8688), pero con un comportamiento distinto: Recall muy alto (0.9445) y Precision mas baja (0.8351). Esto indica que el modelo tiende a capturar la lesión con mayor cobertura (reduciendo falsos negativos) a cambio de incluir pixeles extra (aumentando falsos positivos). Esta diferencia es clave porque anticipa que el "mejor" modelo depende del criterio operacional: minimizar omisiones (Recall) o evitar sobreestimación del area (Precision).

4.3. Análisis de robustez global (degradación ISIC 2018)

Esta sección analiza la resistencia promedio de los modelos en los **58** escenarios de prueba del ISIC 2018, agrupados por bloques de degradación.

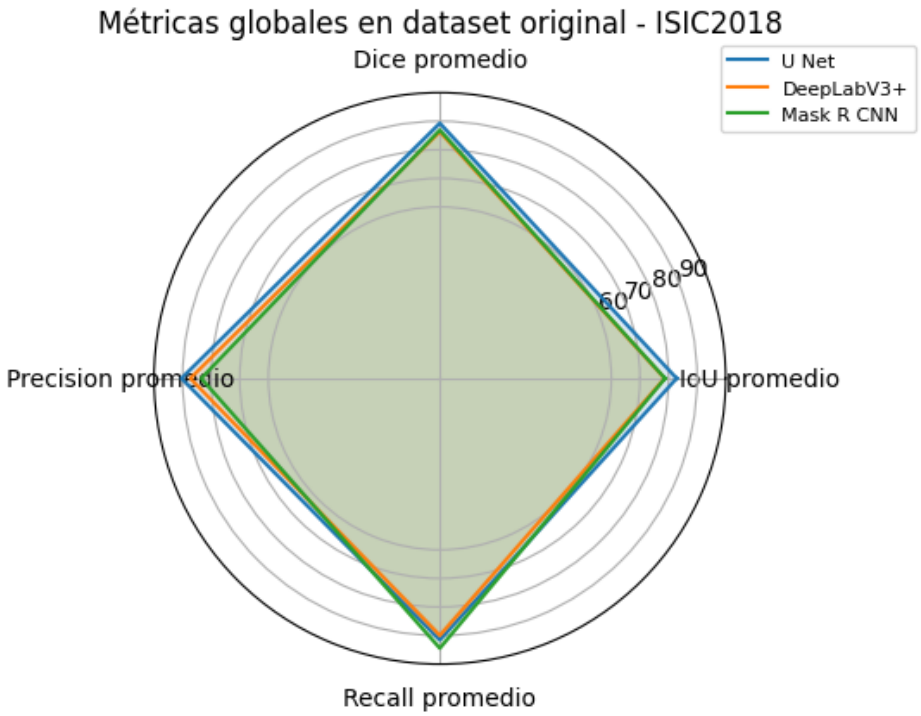


Fig. 4.1: Robustez base (Gráfico Radar).

A partir de la Tabla 4.1 se observan diferencias relevantes entre desempeño promedio y el tipo de error que comete cada arquitectura. U-Net obtiene el mejor desempeño base (IoU = 0.8284, Dice = 0.893), con un balance favorable entre precisión (0.9156) y recall (0.9044), lo que sugiere segmentaciones consistentes sin tendencia marcada a sobre o sub-segmentar.

DeepLabv3+ presenta una disminución respecto a U-Net (IoU = 0.7867, Dice = 0.8631), manteniendo Precision alta (0.8995) pero con menor recall (0.8733). En términos prácticos, este patrón suele asociarse a predicciones mas conservadoras (menos falsos positivos) a costa de perder parte del borde o zonas de baja contrastación.

Mask R-CNN alcanza un IoU similar a DeepLabv3+ (0.788) y Dice (0.8688), pero con un comportamiento distinto: recall muy alto (0.9445) y precisión mas baja (0.8351). Esto indica que el modelo tiende a capturar la lesión con mayor cobertura (reduciendo falsos negativos) a cambio de incluir pixeles extra (aumentando falsos positivos). Esta diferencia es clave porque anticipa que el “mejor” modelo depende del criterio operacional: minimizar omisiones (recall) o evitar sobreestimación del area (precision).

4.3.1. Robustez ante degradaciones simples

Esta subsección evalúa degradaciones simples de manera separada, con el objetivo de identificar cual factor domina la perdida de calidad: (i) reducción de resolución, que elimina detalle de borde y textura; (ii) ruido aditivo (por ejemplo, Gaussiano o sal y pimienta), que reduce la relación señal-ruido y distorsiona gradientes; y (iii) filtros, que pueden suavizar bordes (desenfoque) o alterar la respuesta espacial de la imagen. Los gráficos radar permiten comparar estabilidad promedio entre modelos sin depender de un solo punto de degradación.

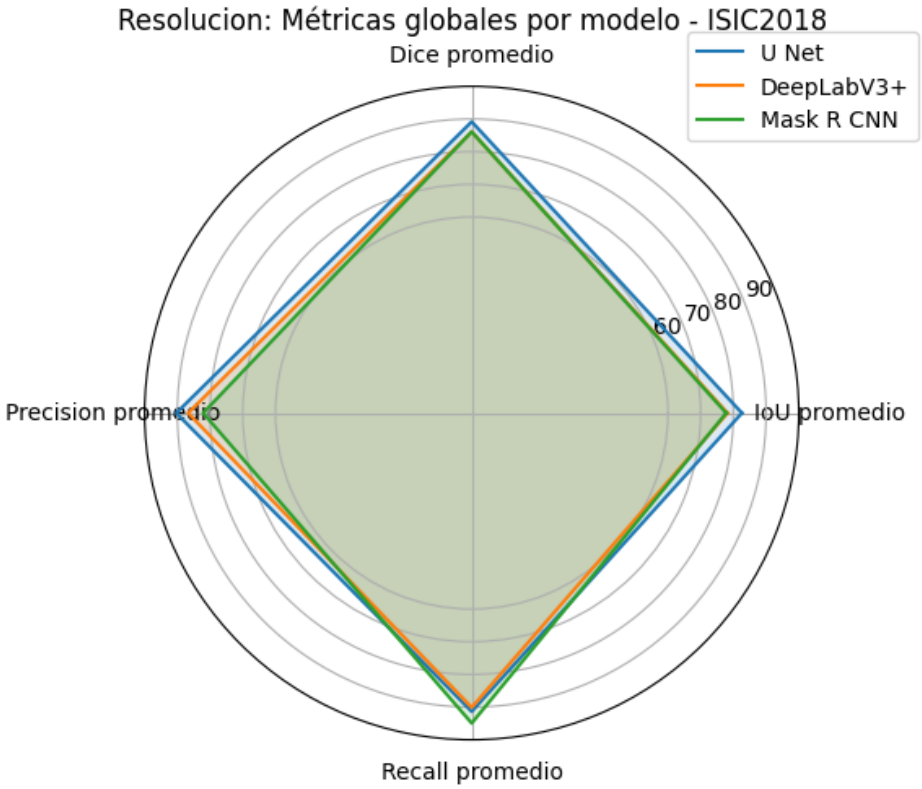


Fig. 4.2: Radar global IoU, Dice, Precision, Recall: Reducción de Resolución.

En la Figura 4.2 se observa que la reducción de resolución impacta principalmente IoU y Dice, debido a la perdida de detalle fino en el contorno de la lesión. Un modelo robusto mantiene su forma radar sin colapsar en estas métricas. En este escenario, el desempeño base no garantiza robustez: el punto crítico es cuanto se contrae el radar al degradar, lo que anticipa la resolución minima recomendable para operar con seguridad.

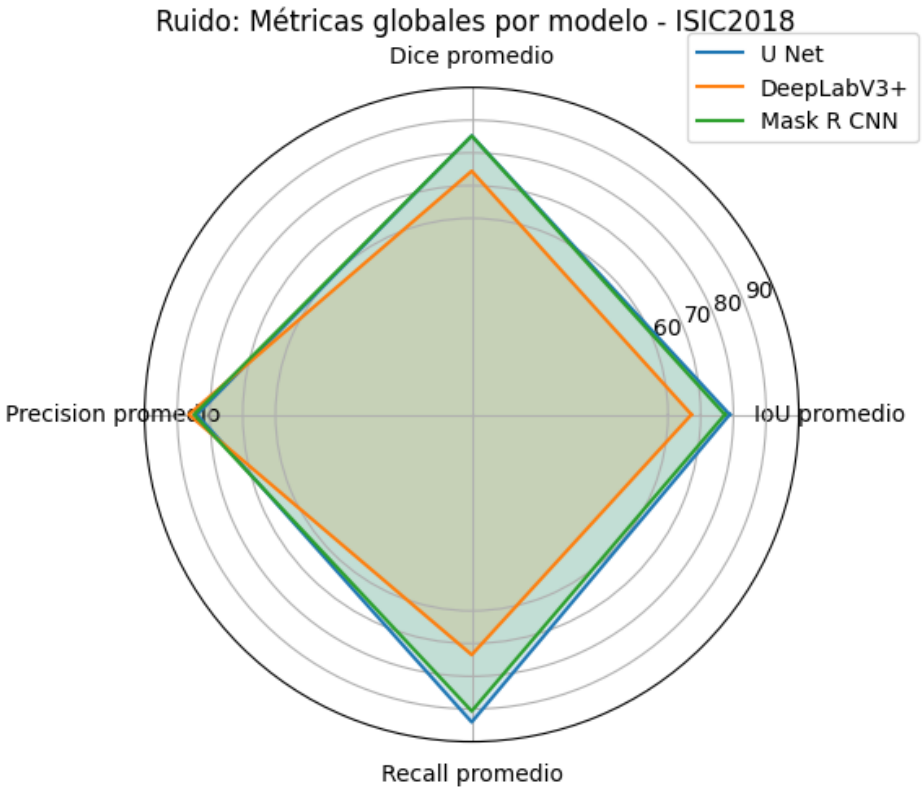


Fig. 4.3: Radar global IoU, Dice, Precision, Recall: Ruido aditivo.

La Figura 4.3 evidencia que el ruido aditivo deteriora el delineado al contaminar textura y borde, afectando la consistencia de la mascara. En general, el ruido tiende a aumentar falsos positivos (baja Precision) o a fragmentar el borde (baja recall), según la arquitectura. La comparación es útil porque distingue sensibilidad estructural: modelos con mejor estabilidad aqui requieren menos preprocesamiento para ser desplegados.

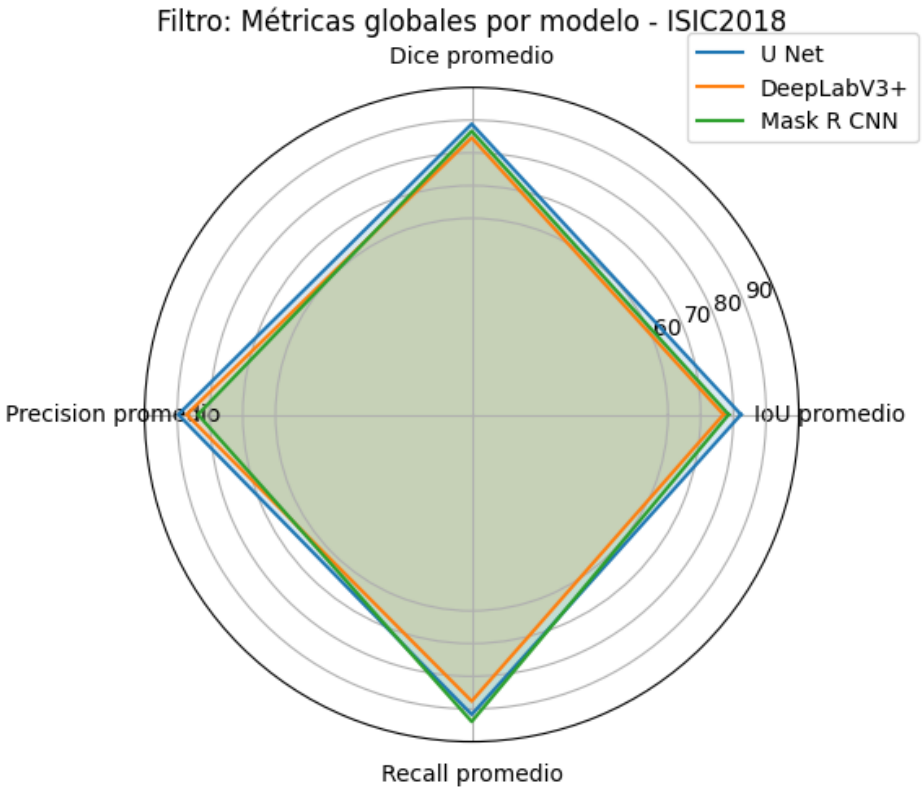


Fig. 4.4: Radar global IoU, Dice, Precision, Recall: Filtro aplicado.

En la Figura 4.4 se aprecia el efecto de filtros sobre la segmentación. Cuando el filtro suaviza, el borde se vuelve ambiguo y suele caer IoU por errores en el contorno; cuando el filtro realza, pueden aparecer artefactos que incrementan falsos positivos. Este bloque permite interpretar si el modelo depende fuertemente de gradientes locales o si logra sostener su predicción con pistas mas globales de forma.

4.3.2. Robustez ante degradaciones combinadas

Las degradaciones combinadas representan escenarios mas cercanos a condiciones reales: una imagen puede venir simultáneamente con baja resolución y ruido, o con ruido mas desenfoque por mala captura. Estas combinaciones no son aditivas de forma lineal: la reducción de resolución puede amplificar la sensibilidad al ruido, y un filtro puede ocultar pistas que el modelo usa para estabilizar el borde. Por ello, este bloque se interpreta buscando colapsos abruptos del perfil radar, mas que disminuciones suaves.

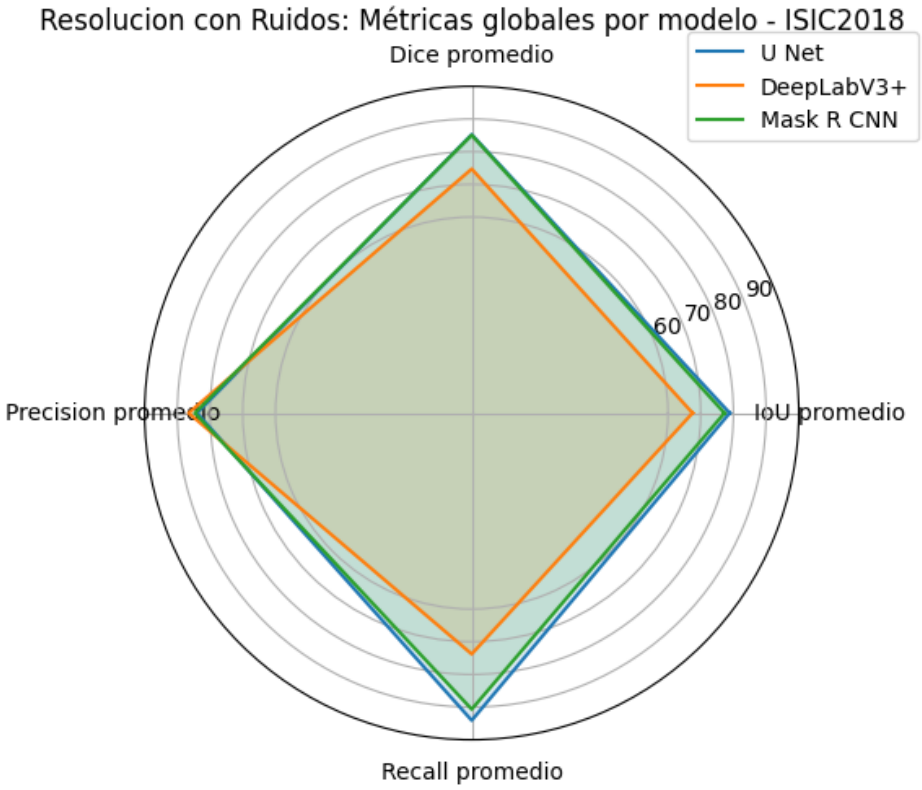


Fig. 4.5: Radar global IoU, Dice, Precision, Recall: Combinación Doble (Resolución + Ruido).

En la Figura 4.5 se observa que la combinación resolución mas ruido es una de las condiciones mas exigentes, porque reduce simultáneamente detalle espacial y relación señal-ruido. Si un modelo mantiene recall alta pero pierde precisión, su falla principal es sobre-segmentación; si cae recall, el riesgo es omitir parte de la lesión, lo cual es operacionalmente mas delicado.

En la Figura 4.6 la degradación dominante suele ser la perdida de borde: la resolución

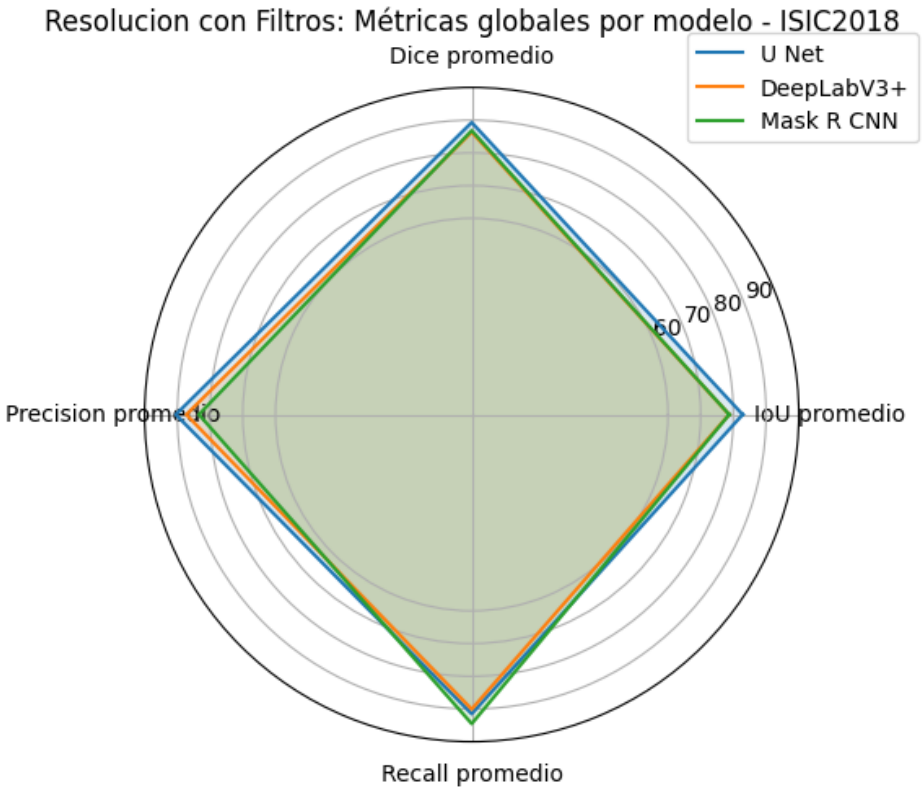


Fig. 4.6: Radar global IoU, Dice, Precision, Recall: Combinación Doble (Resolución + Filtro).

reduce el contorno disponible y el filtro puede suavizarlo aun mas. Un modelo robusto en este caso tiende a conservar Dice relativamente estable, incluso cuando IoU cae moderadamente.

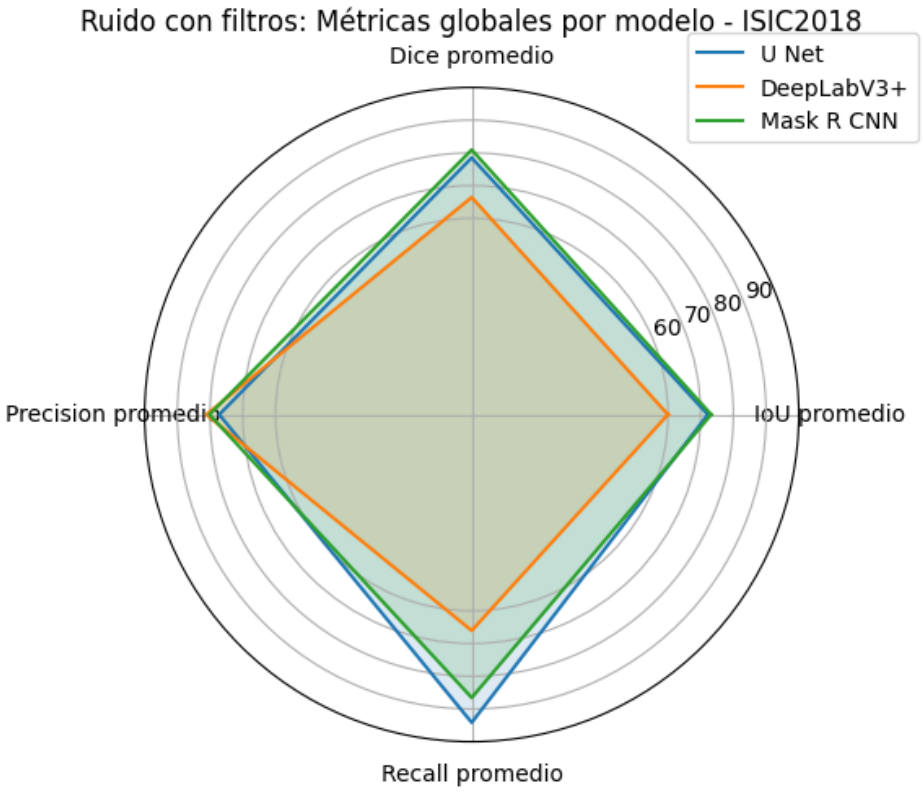


Fig. 4.7: Radar global IoU, Dice, Precision, Recall: Combinación Doble (Resolución + Filtro).

En la Figura 4.7 se aprecia si el filtro mitiga o empeora el efecto del ruido. Si el filtro es suavizante, puede reducir ruido de alta frecuencia pero también borrar borde; por eso es esperable un trade-off entre precision y recall. Este resultado justifica por que el preprocesamiento debe elegirse en funcion del tipo de ruido predominante.

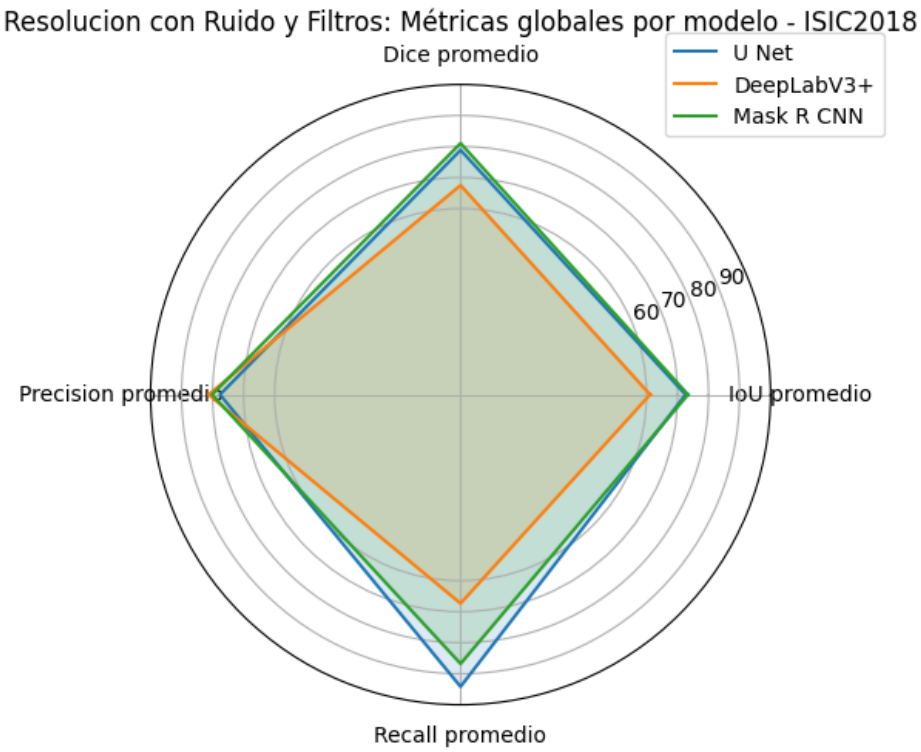


Fig. 4.8: Radar global IoU, Dice, Precision, Recall: Combinación triple (Resolución + Ruido + Filtro).

Finalmente, la Figura 4.8 resume el peor caso. Esta condición es la mas representativa para definir umbrales operacionales, ya que cualquier modelo que sea estable aqui tiende a ser estable en escenarios mas simples. Por lo mismo, el análisis posterior de umbrales se centra en identificar el punto a partir del cual el desempeño deja de ser confiable.

4.3.3. Justificación del umbral operacional

Con el fin de comparar robustez entre modelos y degradaciones, se definió un **umbral operacional** que separa condiciones de funcionamiento aceptable de condiciones donde la segmentación deja de ser confiable. Este umbral no pretende ser un criterio clínico, sino un **estándar mínimo reproducible** para análisis comparativo y para identificar puntos de quiebre bajo degradación.

Consistencia entre IoU y Dice. Se emplean conjuntamente Intersection over Union (IoU) y Dice Coefficient (Dice) porque ambas métricas miden solapamiento, pero

penalizan de forma distinta los errores. IoU es mas estricto ante discrepancias de borde, mientras que Dice suele ser mas alto para una misma predicción. Ambas métricas están relacionadas por:

$$DSC = \frac{2 \cdot IoU}{1 + IoU} \tag{4.1}$$

Por ejemplo, un umbral $IoU = 0{,}75$ equivale aproximadamente a $Dice \approx 0{,}857$. Por esta razón, se adoptó el par **IoU $\geq$ 0.75** y **Dice $\geq$ 0.85**, que es consistente y además levemente conservador para IoU, asegurando que el solapamiento no sea únicamente alto por efecto del tamaño de la lesión.

Umbrales de recall y precisión como lectura del tipo de error. Para complementar IoU y Dice, se incorporan umbrales sobre:

- **Recall $\geq$ 0.85**: prioriza evitar omisiones de la lesión (falsos negativos), especialmente relevante en condiciones adversas donde la segmentación puede fragmentarse o desaparecer.
- **Precisión $\geq$ 0.80**: permite controlar sobre-segmentación (falsos positivos). Se utiliza un umbral ligeramente mas flexible que Recall, ya que en escenarios de degradación severa es esperable cierto exceso de máscara, mientras que la omisión suele ser mas crítica para la cobertura de la lesión.

Criterio práctico basado en la distribución observada. En los gráficos de dispersión IoU vs Dice, la mayor parte de condiciones estables se concentra cerca del cuadrante superior derecho, mientras que las fallas severas aparecen como outliers alejados que cruzan ampliamente los umbrales. Por lo tanto, el umbral operacional cumple una función discriminativa clara: identifica el punto donde el sistema deja de comportarse de manera consistente y pasa a un régimen de error dominante (omisión o sobre-segmentación), lo que se verifica con las barras de Recall y Precision.

En síntesis, el umbral operacional se selecciona por consistencia matemática (IoU-Dice), interpretabilidad del tipo de fallo (recall-precision) y capacidad de separar regímenes de operación en presencia de degradaciones.

4.3.4. Análisis Detallado de Degradaciones Simples

Se presenta los resultados de degradaciones simples, es decir, Reduccion de resolucion, ruido aditivo y filtro aplicado.

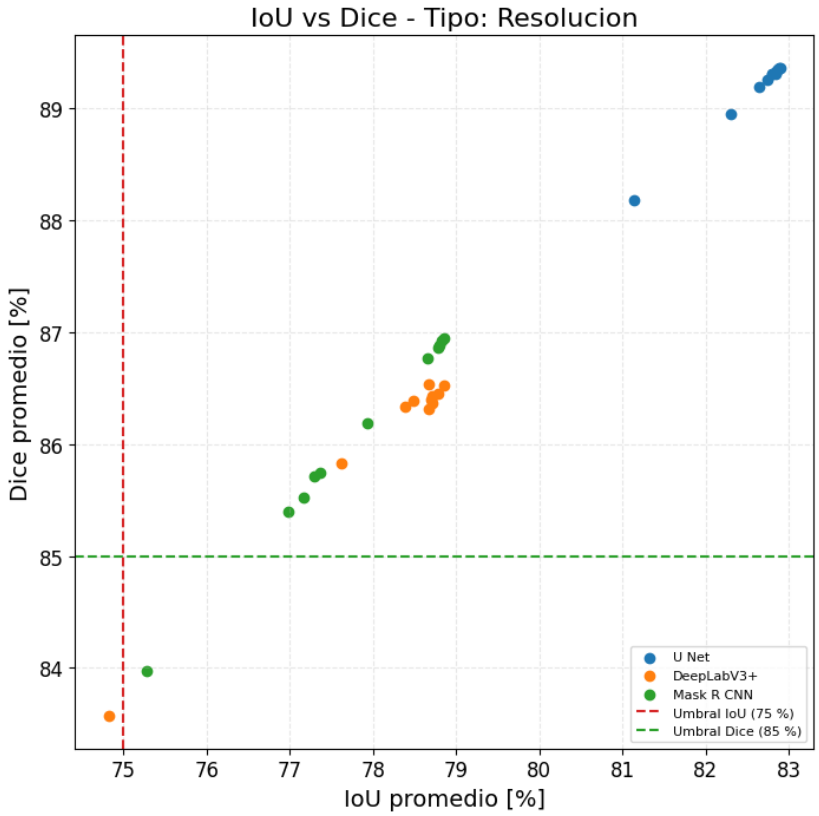


Fig. 4.9: Scatter IoU vs. Dice.

En la figura 4.9 (Intersection over Union, IoU, vs Dice Similarity Coefficient, Dice) se observa una degradación progresiva a medida que disminuye la resolución. U-Net se mantiene sistemáticamente en la zona superior derecha, con valores cercanos a IoU $\approx$ 82–83 % y Dice $\approx$ 89 %, lo que indica alta estabilidad ante cambios de escala. En contraste, DeepLabV3+ y Mask Region-based Convolutional Neural Network (Mask R-CNN) se concentran en un rango intermedio (IoU $\approx$ 77–79 %, DSC $\approx$ 85,4–87 %), mostrando mayor sensibilidad a la pérdida de detalle fino en el borde.

El caso mas crítico ocurre en el nivel mas bajo (3 %), donde DeepLabV3+ cae bajo ambos umbrales (IoU $\approx$ 74,8 % y Dice $\approx$ 83,6 %) y Mask R-CNN queda marginalmente sobre el umbral de IoU (IoU $\approx$ 75,2 %) pero bajo el umbral de Dice (Dice $\approx$ 84,0 %).

Este patrón es consistente con errores dominados por contorno: la baja resolución reduce información espacial y desplaza el borde estimado varios píxeles, lo que penaliza IoU y también Dice cuando la desviación se vuelve sistemática.

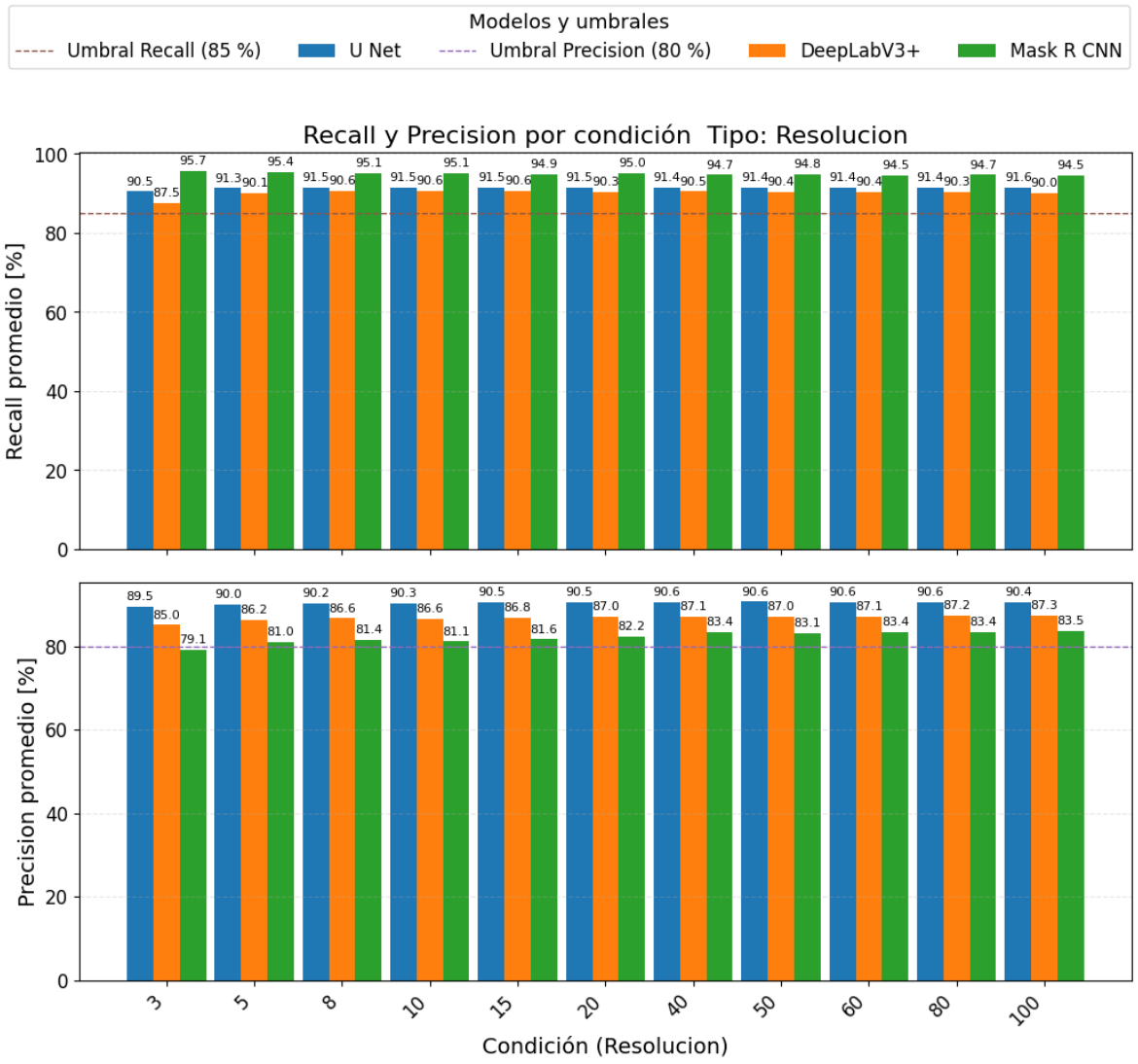


Fig. 4.10: Recall y Precisión por condición.

La figura 4.10 confirma que la reducción de resolución no produce una pérdida marcada de detección de la lesión: el recall se mantiene por sobre el umbral de 85 % en todos los niveles evaluados para los tres modelos (por ejemplo, DeepLabV3+ se mantiene entre 87,5 y 90,6 %, y Mask R-CNN entre 94,5 y 95,7 %). Esto sugiere que el problema principal no es omitir la lesión, sino delimitarla con precisión.

La diferencia aparece en precisión. U-Net conserva precisión alta y estable (cercana

a 89,5–90,6 %), y DeepLabV3+ también se mantiene sobre 85 %. Mask R-CNN, en cambio, presenta su mínimo en 3 % (precisión 79,1 %), bajo el umbral de 80 %. Esto indica tendencia a sobre-segmentación en condiciones de resolución extrema: el modelo mantiene recall alto, pero incluye tejido sano, lo que explica por qué en el scatter su Dice cae bajo el umbral aun cuando IoU no colapsa completamente. En términos operacionales, el nivel de 3 % se vuelve condición límite, y niveles desde 5 % en adelante muestran un comportamiento mas consistente.

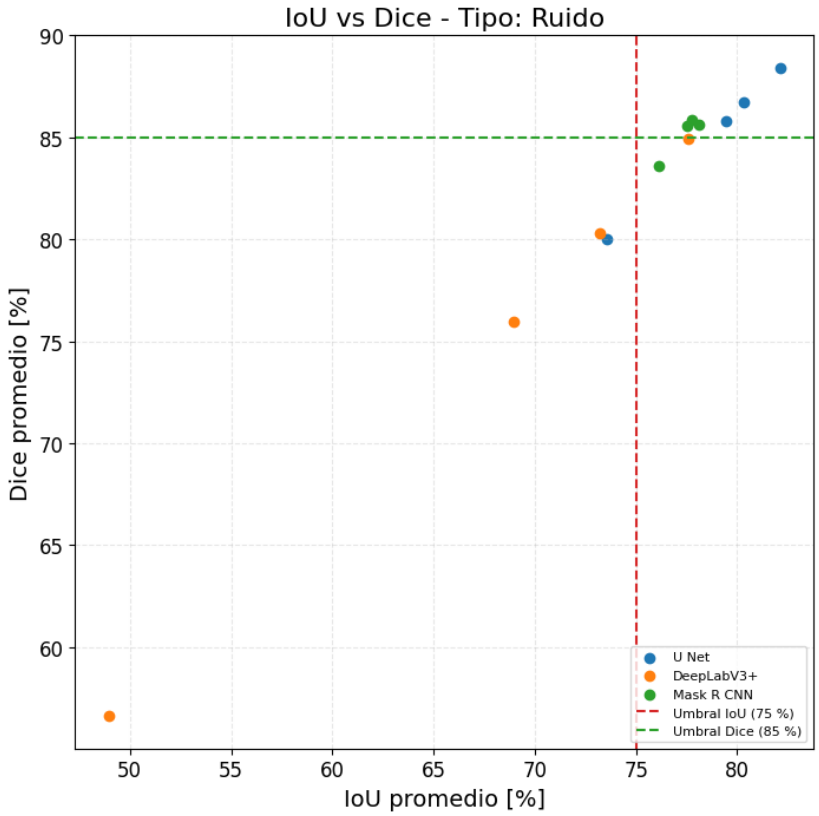


Fig. 4.11: Scatter IoU vs. Dice.

En la figura 4.11 (Intersection over Union, IoU, vs Dice Similarity Coefficient, Dice) el ruido es la degradación que genera mayor dispersión y los outliers mas severos, indicando sensibilidad a la relación señal-ruido y a la aparición de texturas espurias. Para DeepLabV3+ se observa un colapso fuerte en el régimen mas severo, con un punto cercano a IoU $\approx$ 49 % y Dice $\approx$ 56,7 %, muy por debajo de los umbrales (IoU 75 %, Dice 85 %). Además, existe un punto intermedio alrededor de IoU $\approx$ 69 % y Dice $\approx$ 76 %, que confirma que el deterioro no es gradual, sino que presenta quiebres al aumentar la intensidad del ruido.

U-Net muestra alta estabilidad en condiciones moderadas (puntos en IoU ≈ 79–82 % y Dice ≈ 86–88,5 %), pero también presenta una caída bajo umbral en la condición mas exigente (aprox. IoU ≈ 73,6 %, Dice ≈ 80 %), lo que evidencia que, sin control de calidad o preprocesamiento, el ruido puede llevar a un régimen no confiable incluso en el modelo mas fuerte. Mask R-CNN se mantiene mas concentrado cerca de los umbrales, con varios puntos en IoU ≈ 76–78 % y Dice ≈ 83,7–86 %; en este caso, el efecto dominante se refleja en Dice cercano o levemente bajo 85 %, indicando degradación por borde y fragmentación de la máscara mas que una falla total.

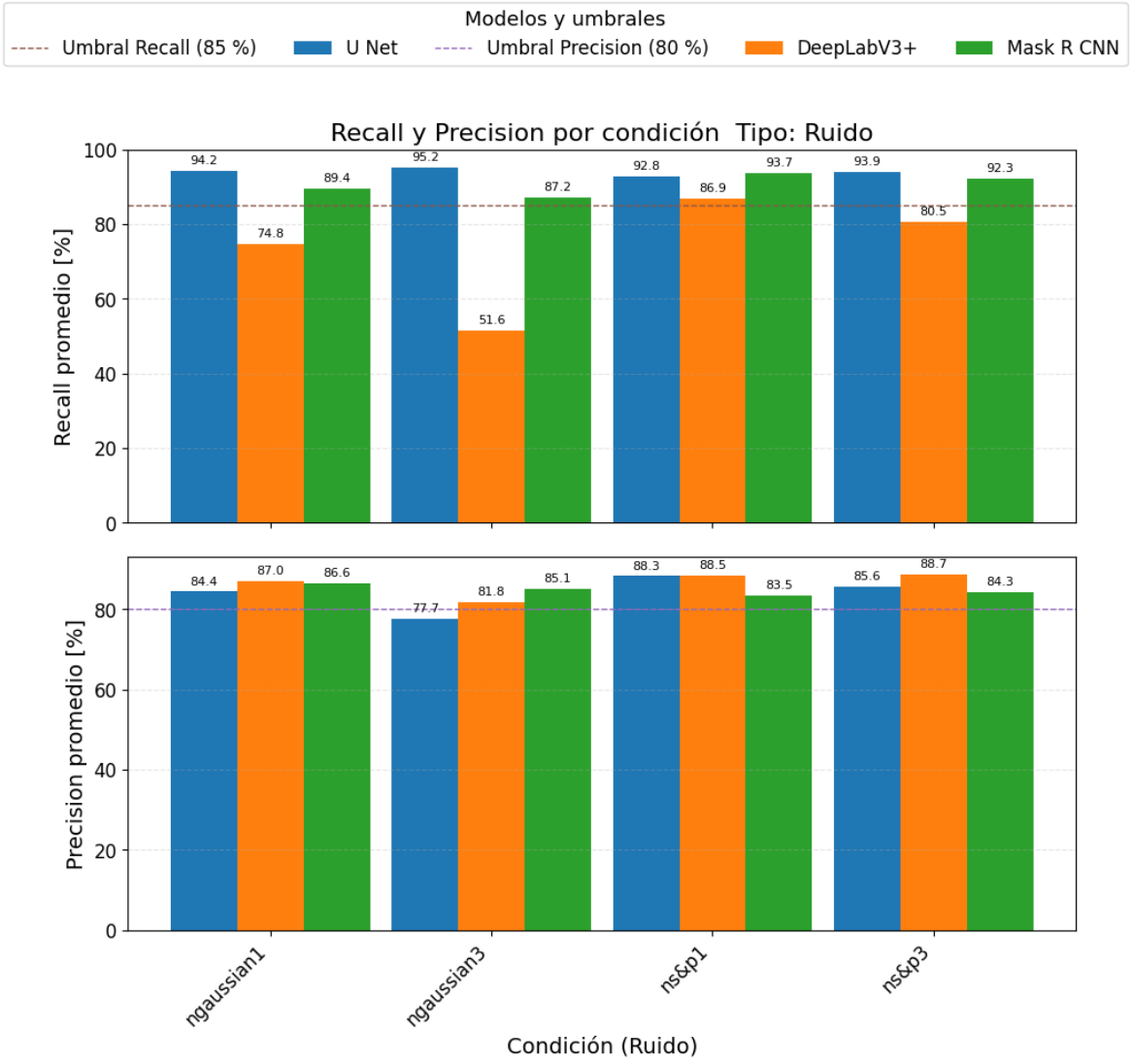


Fig. 4.12: Recall y precisión por condición.

La figura 4.12 permite identificar el tipo de error. En DeepLabV3+ el recall cae

drásticamente con ruido Gaussiano: 74,8 % (ruido gaussiano con varianza 0,01) y 51,6 % (ruido gaussiano con varianza 0,03), muy bajo el umbral de 85 %, mientras su precision se mantiene sobre 80 % (por ejemplo, 81,8 % en ruido gaussiano con varianza 0,03). Este patrón describe una segmentación conservadora: el modelo evita falsos positivos pero omite gran parte de la lesión, lo cual es operacionalmente crítico si el objetivo es no subestimar el área lesionada.

En U-Net ocurre el fenómeno inverso bajo el ruido mas severo: el Recall se mantiene alto (por ejemplo 95,2 % en ruido gaussiano con varianza 0,03), pero la precisión baja a 77,7 %, bajo el umbral de 80 %. Esto sugiere sobre-segmentación inducida por ruido: el modelo cubre la lesión pero incorpora tejido sano, lo que reduce IoU y Dice por exceso de máscara. Mask R-CNN mantiene recall alto en todas las condiciones (mínimo 87,2 % en ruido gaussiano con varianza 0,03) y precisión estable sobre 83 %, mostrando el perfil mas robusto frente a ruido. En conjunto, estos resultados justifican que el ruido, especialmente el gaussiano, requiere una etapa previa de mitigación o un control de calidad de entrada para evitar operar en el régimen de colapso observado.

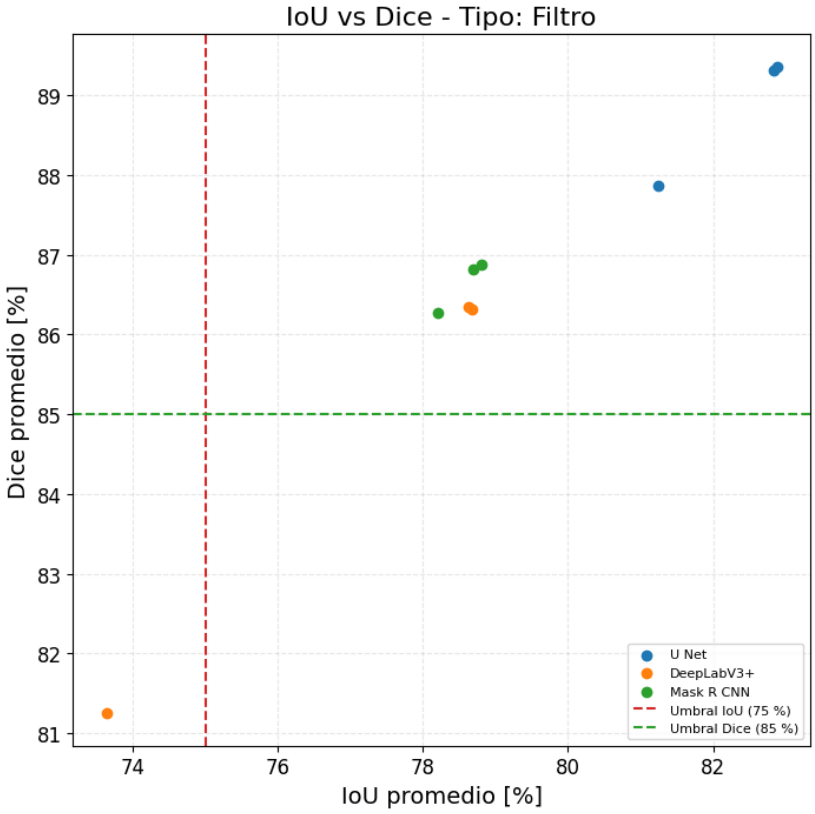


Fig. 4.13: Scatter IoU vs. Dice.

En la figura 4.13 (Intersection over Union, IoU, vs Dice Similarity Coefficient, Dice) se aprecia que la familia de filtros produce un impacto menor que el ruido y, en general, los puntos se mantienen sobre los umbrales (IoU 75 %, Dice 85 %). U-Net conserva el mejor posicionamiento, con puntos en IoU $\approx$ 81–83 % y Dice $\approx$ 87,9–89,3 %, reflejando alta estabilidad ante modificaciones de nitidez y suavizado.

Mask R-CNN y DeepLabV3+ se agrupan cerca de IoU $\approx$ 78,2–78,8 % y Dice $\approx$ 86,3–86,9 %, lo que sugiere que los filtros alteran parcialmente el borde pero sin provocar un quiebre generalizado. La excepción relevante es un outlier de DeepLabV3+ en IoU $\approx$ 73,7 % y Dice $\approx$ 81,2 %, bajo ambos umbrales, indicando que una condición de filtrado específica puede inducir un error sistemático de contorno o una pérdida de contraste útil para el modelo.

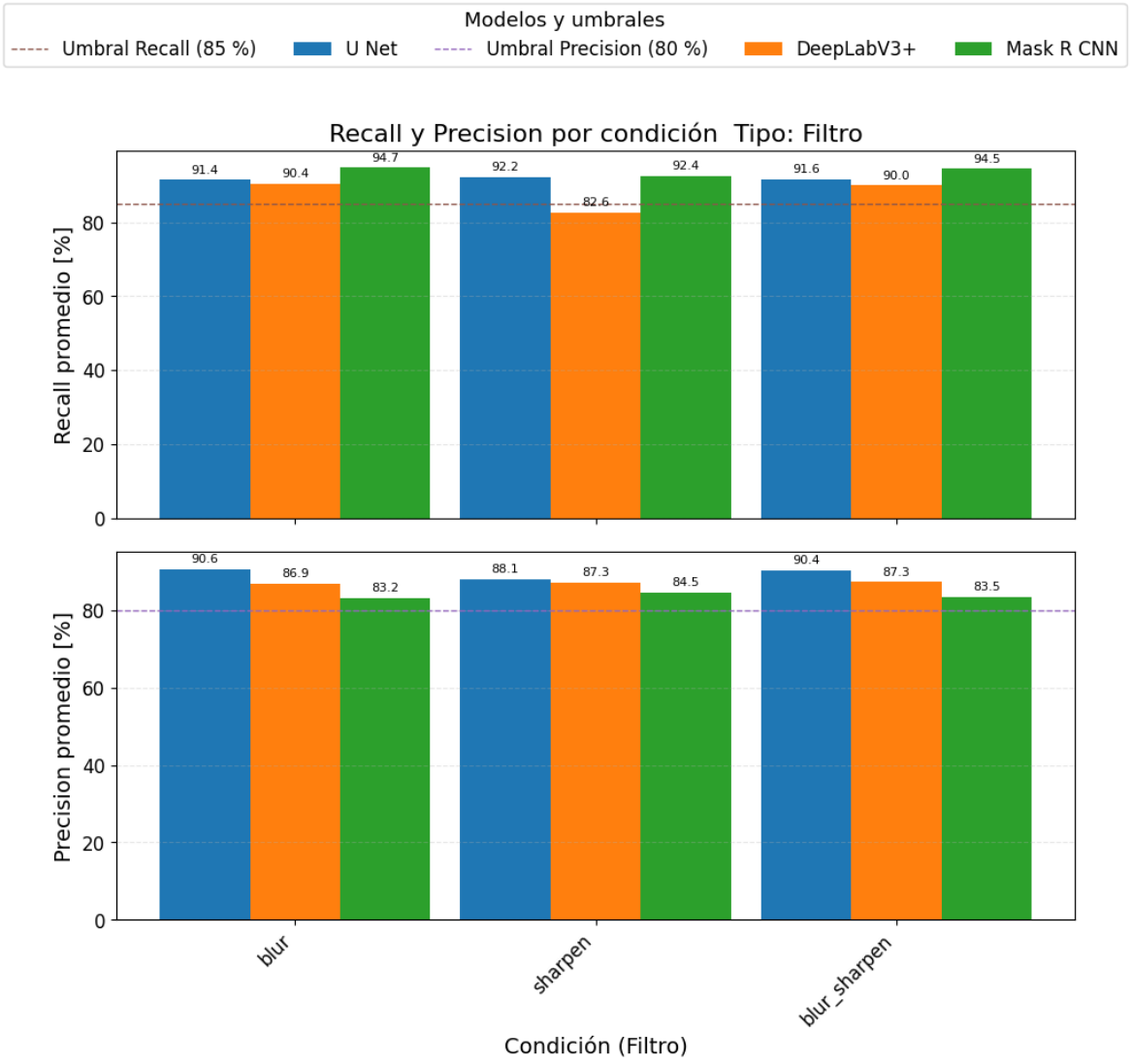


Fig. 4.14: Recall y precisión por condición.

La figura 4.14 confirma que, para U-Net y Mask R-CNN, el filtrado no compromete la cobertura de la lesión: el recall se mantiene sobre 91 % en U-Net y sobre 92 % en Mask R-CNN en las tres condiciones. DeepLabV3+ presenta un comportamiento mas sensible en la condición de sharpen, donde su recall baja a 82,6 % (bajo el umbral 85 %), lo que implica omisión parcial de la lesión. Esta caída es consistente con la idea de que ciertos realces pueden introducir artefactos o modificar gradientes de borde de forma que el modelo reduzca su área predicha.

En precisión, los tres modelos se mantienen sobre el umbral de 80 % en todas las condiciones. U-Net registra los valores mas altos (por ejemplo 90,6 % en blur),

DeepLabV3+ se mantiene alrededor de 86,9–87,3 %, y Mask R-CNN presenta los valores mas bajos pero aceptables (83,2–84,5 %), coherente con su tendencia general a segmentar de forma mas inclusiva. En términos operacionales, el filtrado es una degradación menos crítica que el ruido, pero conviene evitar configuraciones que reduzcan recall en DeepLabV3+ si el criterio prioritario es minimizar omisiones.

4.3.5. Análisis Detallado de Degradaciones Combinadas

Se presenta los resultados de las combinaciones dobles, es decir, reducción de resolución con ruido aditivo, reducción de resolución con filtro aplicado y ruido aditivo con filtro aplicado.

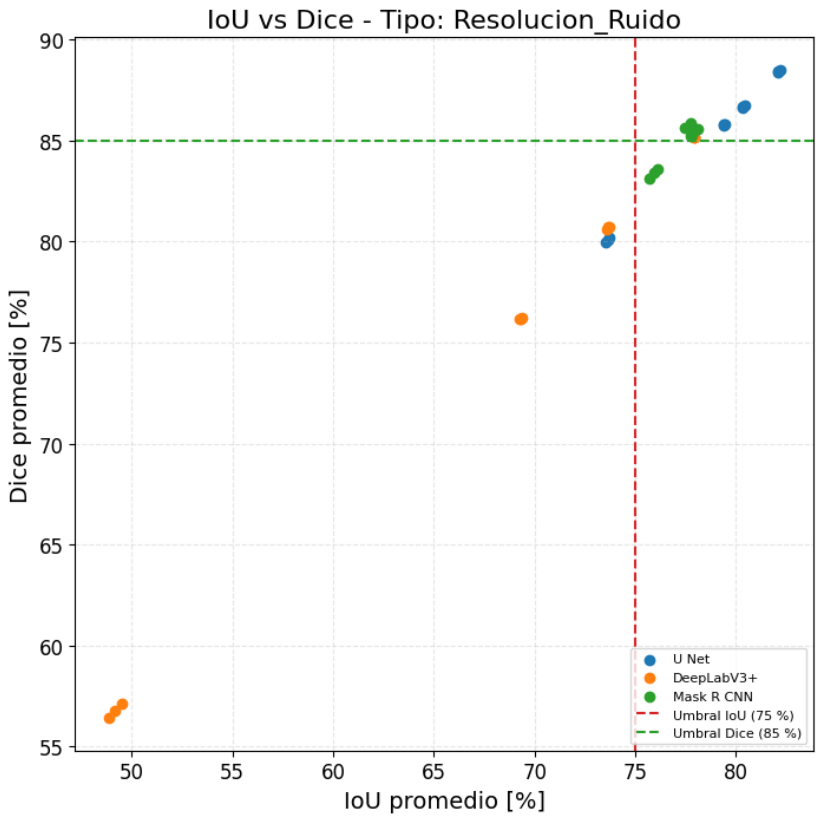


Fig. 4.15: Scatter IoU vs. Dice.

En la figura 4.15 (IoU vs Dice) se observa que la degradación dominante sigue siendo el ruido, y la reducción de resolución no compensa su efecto en el régimen severo. Se distinguen nuevamente dos comportamientos: un cluster cercano a condiciones

aceptables (especialmente con sal y pimienta) y puntos de colapso asociados al ruido Gaussiano de mayor intensidad.

Los outliers mas severos corresponden a DeepLabV3+ bajo ruido Gaussiano intenso, donde los puntos caen muy por debajo de los umbrales (IoU alrededor de 49-50 % y Dice alrededor de 56-57 %). En U-Net y Mask R-CNN la caída es menos extrema, pero se observa desplazamiento hacia abajo del umbral en las condiciones mas exigentes, reflejando pérdida de precisión del contorno y degradación del solapamiento.

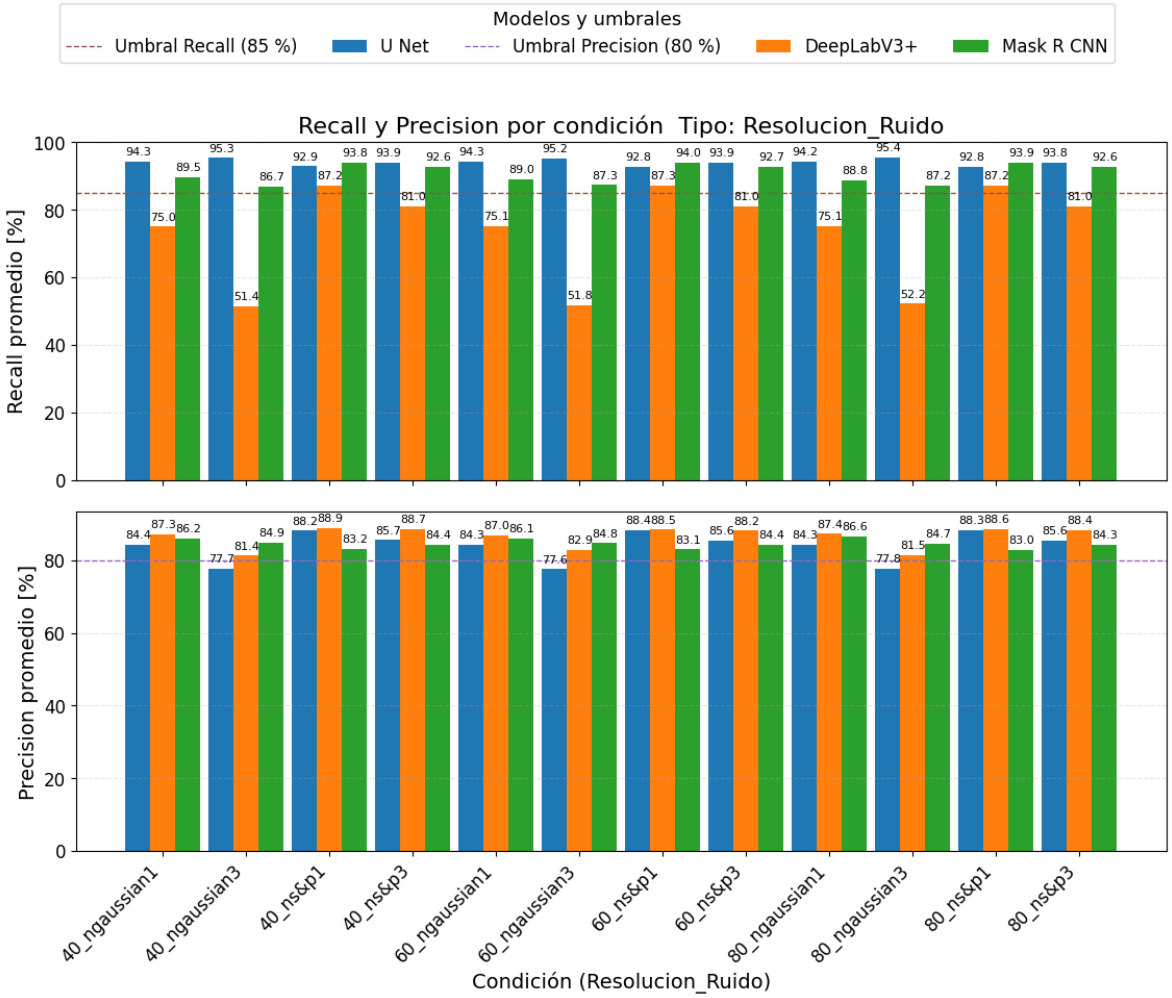


Fig. 4.16: Recall y precisión por condición.

La figura 4.16 muestra que el modo de falla depende del modelo. En DeepLabV3+, el ruido gaussiano reduce principalmente el Recall: por ejemplo, para resolucion de 40 % ruido gaussiano con varianza 0,03 el recall es 51.4 %, para resolución de 60 % con

ruido gaussiano con varianza 0,03 es 51.8 % y para resolución de 80 % ruido gaussiano con varianza 0,03 es 52.2 %. Incluso con ruido gaussiano con varianza 0,01, el recall se mantiene alrededor de 75 %, bajo el umbral de 85 %. Esto confirma una falla por omisión consistente: el modelo se vuelve conservador y deja de cubrir parte importante de la lesión.

En U-Net ocurre lo inverso bajo ruido Gaussiano severo: el recall se mantiene alto (por ejemplo 95.3, 95.2 y 95.4 % en resolución de 40 % con ruido gaussiano con varianza 0,03, resolucion 60 % con ruido gaussiano con varianza 0,03 y resolucion 80 % ruido gaussiano con varianza 0,03), pero la Precision cae bajo el umbral (77.7, 77.6 y 77.8 %). Esto describe sobre segmentación inducida por ruido, lo que deteriora IoU y DSC por exceso de máscara.

Mask R-CNN mantiene un perfil mas balanceado: aunque su recall disminuye en condiciones severas (por ejemplo 86.7 % en resolución de 40 % con ruido gaussiano con varianza 0,03) y su precisión se mantiene sobre 80 %, evita el colapso observado en DeepLabV3+ y el desplome de precisión observado en U-Net. Para ruido sal y pimienta, los tres modelos mantienen métricas cercanas o sobre umbral, indicando que el problema crítico del bloque es el ruido gaussiano.

En síntesis, en resolución + ruido el factor crítico es el ruido Gaussiano: DeepLabV3+ falla por omisión (recall bajo), U-Net falla por sobre segmentación (precisión baja) y Mask R-CNN es el mas estable de los tres. Esto justifica que los umbrales operacionales se definan principalmente en función de ruido, mas que de resolución, cuando ambos están presentes.

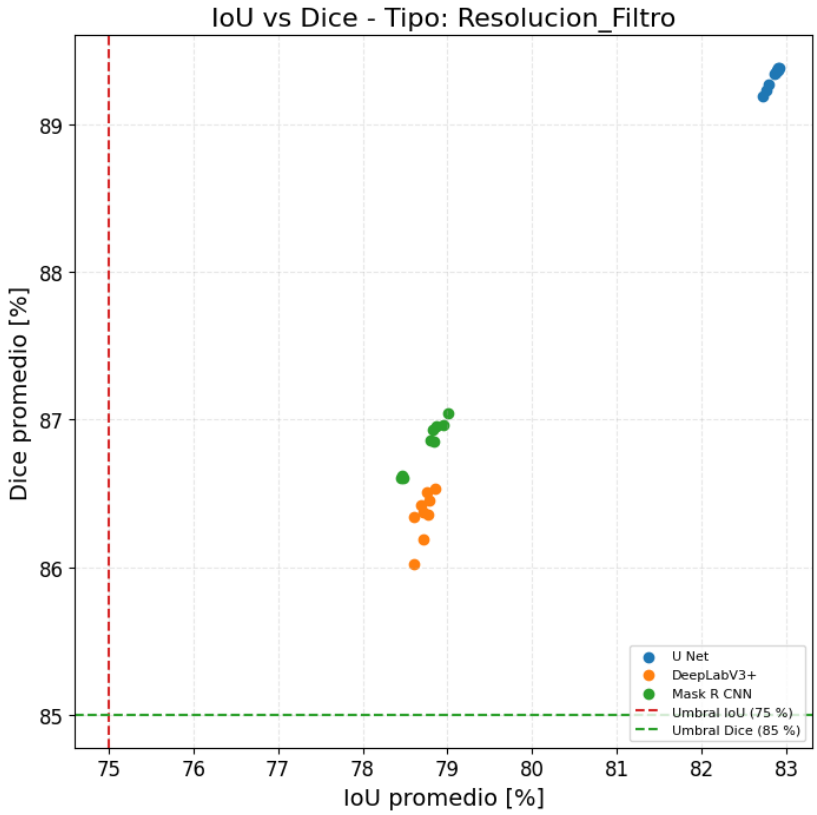


Fig. 4.17: Scatter IoU vs. Dice.

En la figura 4.17 (Intersection over Union, IoU, vs Dice Similarity Coefficient, Dice) se observa que la combinación resolución + filtro no genera un colapso del desempeño dentro del rango evaluado (40, 60 y 80 % con blur, sharpen y blur+sharpen). Los puntos de U-Net se mantienen agrupados en la zona superior derecha (IoU cercano a 82.7-83.0 % y Dice cercano a 89.2-89.4 %), evidenciando alta estabilidad frente a cambios de nitidez y suavizado, incluso cuando la resolución se reduce.

DeepLabV3+ y Mask Region based Convolutional Neural Network (Mask R-CNN) se concentran en un cluster estable alrededor de IoU cercano a 78.5-79.0 % y Dice cercano a 86.0-87.1 %. Lo relevante es que no se observan puntos cruzando los umbrales operacionales (IoU 75 %, Dice 85 %), por lo que el filtrado, en este rango, actúa mas como perturbación moderada del contorno que como una degradación crítica.

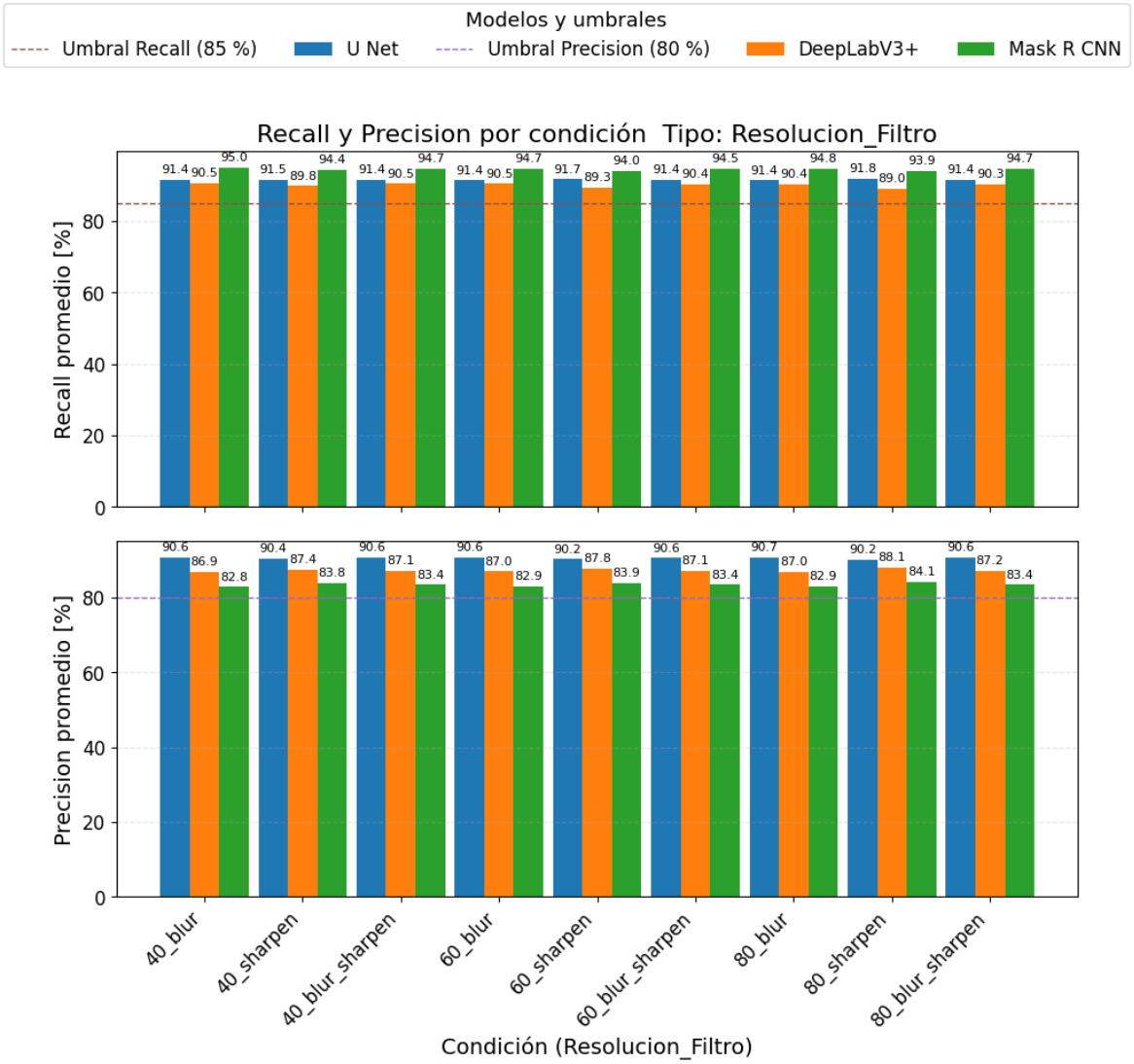


Fig. 4.18: Recall y precisión por condición.

En la figura 4.18 se observa que Mask Region based Convolutional Neural Network (Mask R-CNN) presenta un recall mayor a U-Net y DeepLabV3+ alrededor de 93.9-95.0 % pero una precision alrededor de 82.8- 84.1 % el cual es menor a los otros modelos en donde U-Net presenta una mayor precision, por lo tanto, en las combinaciones de resolución + filtro el modelo de Mask R-CNN presenta mas casos de FP (falsos positivos) que los otros modelos.

En términos operacionales, resolución + filtro (40-80 %) no define el peor caso del sistema: es una combinación relativamente segura comparada con escenarios que incluyen ruido, y sirve mas para medir estabilidad del contorno que para fijar limites

críticos.

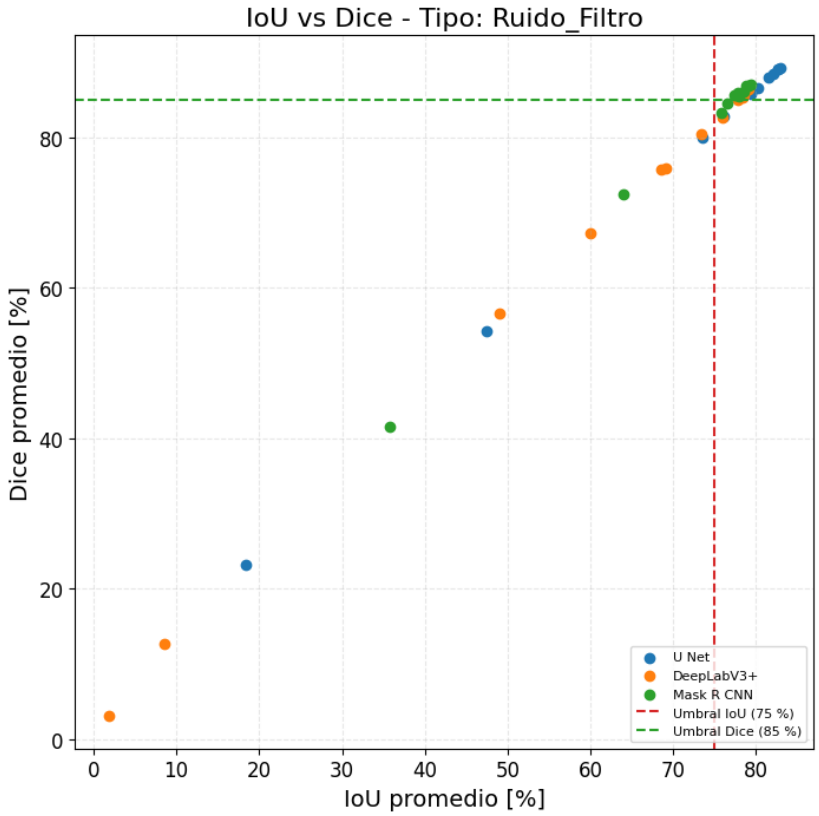


Fig. 4.19: Scatter IoU vs. Dice.

En la figura 4.19 (IoU vs Dice) se aprecia el caso mas inestable entre las combinaciones dobles. La nube presenta dos regímenes claramente separados: (a) un grupo cercano a la zona aceptable (IoU alrededor de 76-83 %, DSC alrededor de 82-88 %), y (b) outliers severos con IoU y Dice extremadamente bajos, lo que indica fallas catastróficas de segmentación bajo ciertas combinaciones específicas.

En particular, las combinaciones de ruido Gaussiano con sharpen generan los puntos mas alejados del umbral, evidenciando que el realce puede amplificar el ruido y distorsionar los patrones de borde usados por los modelos. En contraste, cuando el ruido se combina con blur, el desempeño tiende a concentrarse mas cerca del cluster superior, sugiriendo que el suavizado actúa como mitigación parcial del ruido en varios casos.

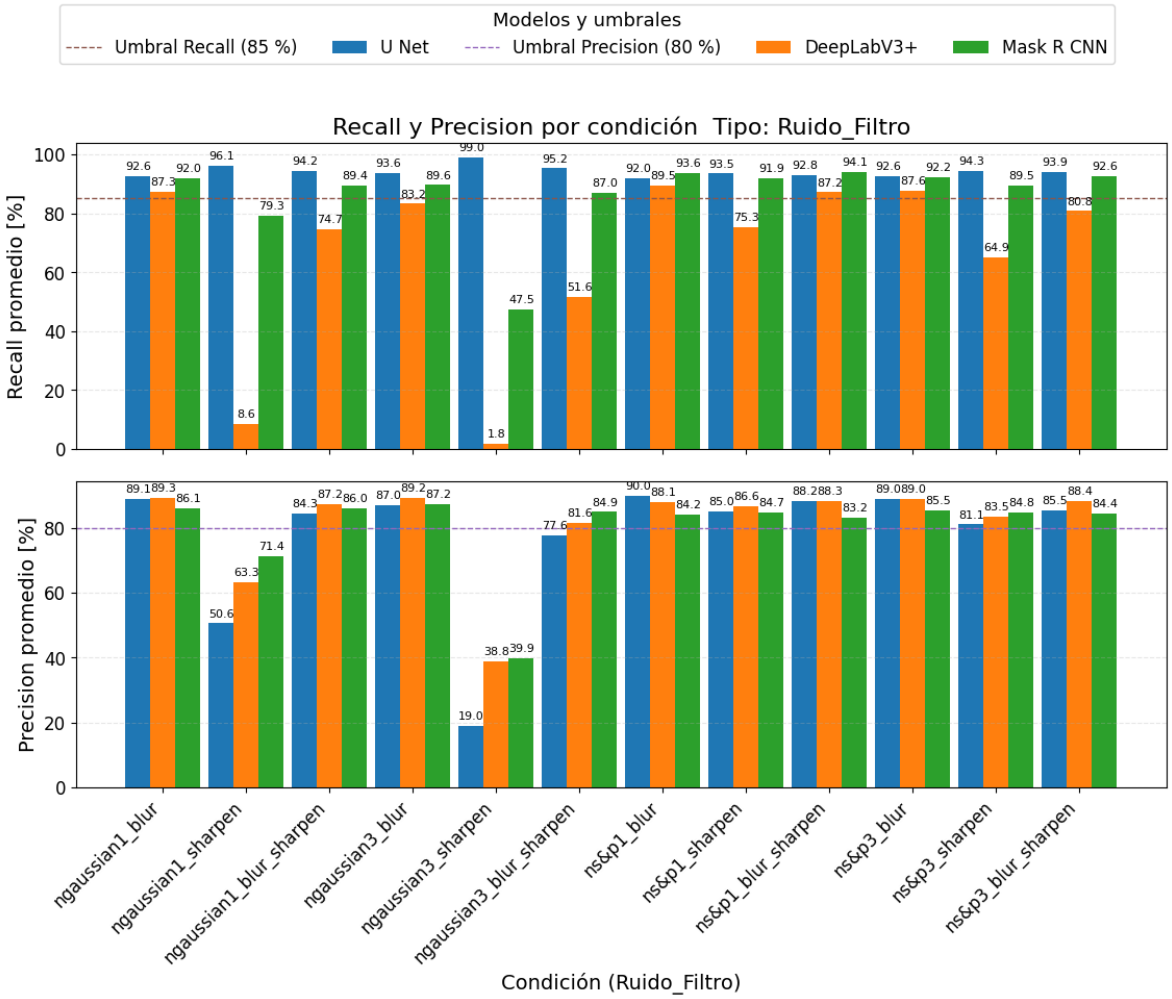


Fig. 4.20: Recall y precisión por condición.

La figura 4.20 permite identificar con claridad el tipo de falla en las condiciones críticas. Para DeepLabV3+ se observa un colapso extremo de recall bajo ruido gaussiano con sharpen: por ejemplo, en ruido gaussiano con varianza 0,01 y filtro sharpen el recall cae a 8.6 %, y en ruido gaussiano con varianza 0,03 y filtro sharpen cae a 1.8 %, muy por debajo del umbral de 85 %. Esto describe una falla por omisión: el modelo prácticamente deja de segmentar la lesión.

En U-Net ocurre un patrón distinto en el régimen mas severo: en ruido gaussiano con varianza 0,03 con filtro sharpen el recall se mantiene alto (99.0 %), pero la precisión cae fuertemente (19.0 %), lo que indica sobre segmentación extrema (predice gran parte de la imagen como lesión). Este comportamiento explica por qué aparecen outliers de IoU y Dice muy bajos aun cuando el recall no disminuye. Mask R-CNN muestra mayor

resistencia relativa, pero aun así cae en condiciones críticas (por ejemplo, recall 47.5 % en ruido gaussiano con varianza 0,03 con filtro sharpen), confirmando que la combinación gaussiano + sharpen es estructuralmente adversa para las tres arquitecturas.

Para ruido sal y pimienta, el comportamiento es mas estable: la mayoría de las condiciones mantienen recall alto en U-Net y Mask R-CNN, mientras que DeepLabV3+ tiende a perder recall en escenarios mas agresivos (por ejemplo, ruido sal y pimienta con densidad 0,03 con filtro sharpen con recall 64.9 %), consistente con omisión parcial.

Operacionalmente, el peor caso del bloque Ruido + Filtro se concentra en ruido Gaussiano combinado con sharpen. Si el sistema debe operar en presencia de ruido, conviene evitar realce agresivo previo, y es preferible un suavizado controlado (blur) o una etapa de denoising antes de segmentar.

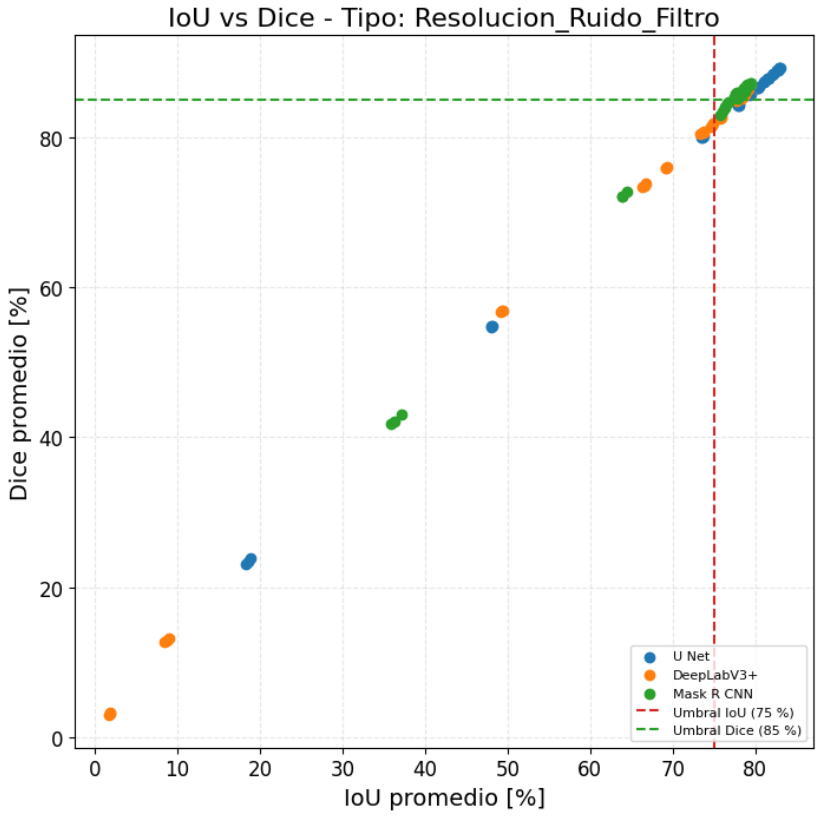


Fig. 4.21: Scatter IoU vs. Dice.

En la figura 4.21 (Intersection over Union, IoU, vs Dice Similarity Coefficient, Dice) se observa el escenario mas exigente del estudio, donde aparecen claramente dos regímenes: (i) un grupo estable cercano a la zona superior derecha y (ii) un conjunto de

outliers severos con IoU y Dice extremadamente bajos. El grupo estable se concentra aproximadamente en IoU $\approx$ 78–83 % y Dice $\approx$ 84–89 %, lo que corresponde a condiciones donde, a pesar de la degradación, los modelos siguen segmentando con solapamiento razonable.

Sin embargo, el régimen crítico evidencia colapsos que cruzan ampliamente los umbrales operacionales (IoU 75 % y Dice 85 %). Estos outliers no representan una degradación leve del contorno, sino fallas estructurales de la segmentación. En particular, la presencia de puntos con IoU menor a 50 % y Dice menor a 60 % indica que, bajo ciertas combinaciones, el modelo deja de mantener coherencia espacial con la máscara real, ya sea por omisión casi completa o por sobre segmentación extrema. Además, se aprecia que la reducción de resolución (40, 60 y 80 %) no evita el colapso cuando el tipo de ruido y el filtro son adversos, lo que sugiere que en este bloque el factor dominante es la interacción entre ruido y filtrado.

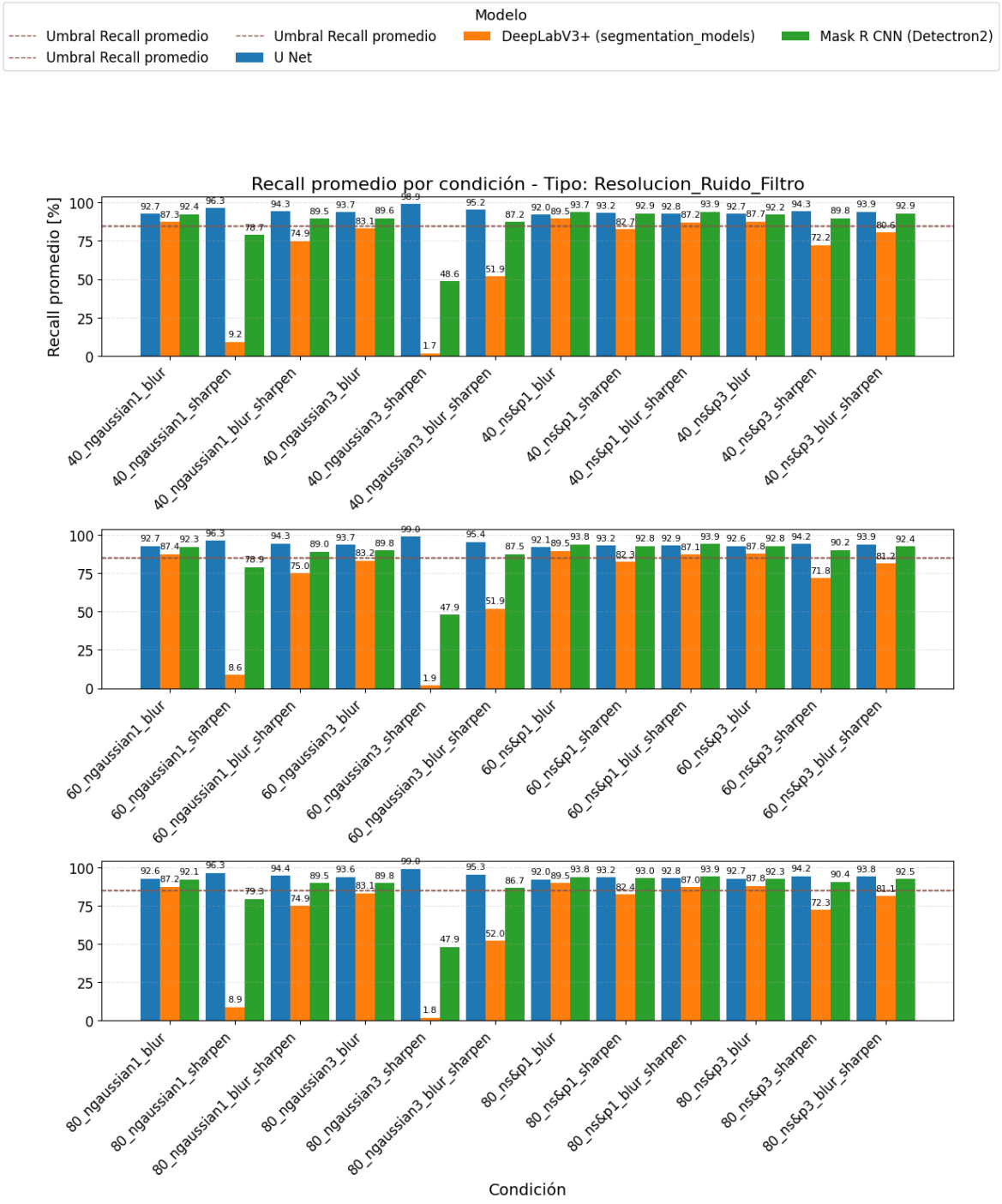


Fig. 4.22: Recall por condición.

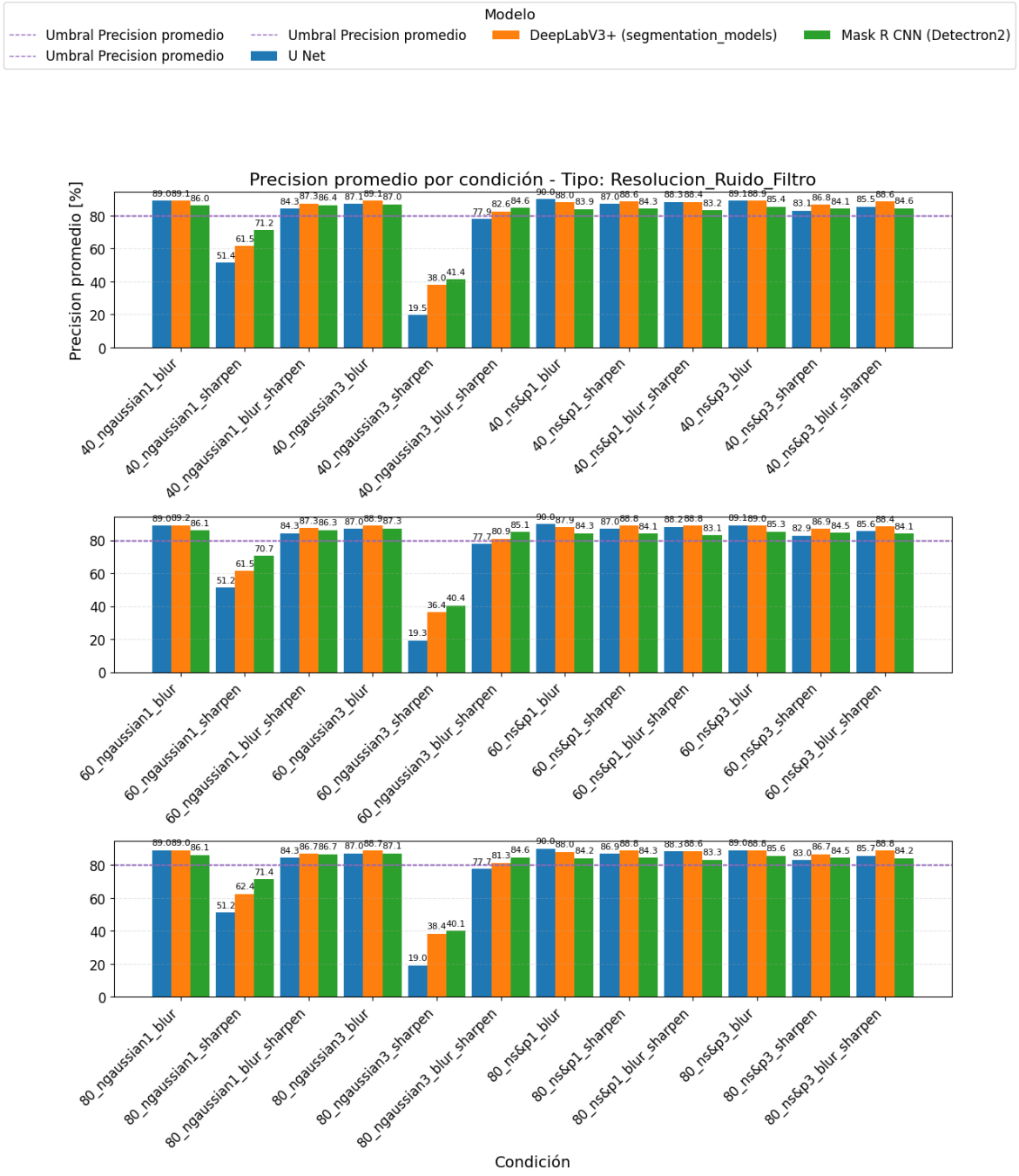


Fig. 4.23: Precisión por condición.

Las figuras 4.23 y 4.22 permiten identificar el modo de falla dominante por arquitectura, usando precisión (precisión) y exhaustividad (recall). El patrón mas crítico ocurre cuando el ruido Gaussiano se combina con sharpen. En DeepLabV3+ el recall colapsa casi a cero en esa condición, por ejemplo: 9,2 % (resolucion de 40 % con ruido gaussiano

de varianza 0,01 y con filtro sharpen), 8,6 % (resolucion de 60 % con ruido gaussiano con varianza de 0,01 y con filtro sharpen) y 8,9 % (resolucion de 80 % ngaussian con varianza 0,01 con filtro sharpen). Para ruido gaussiano mas severo el efecto es aun mayor, con recall alrededor de 1,7–1,9 % en ruido gaussiano con varianza 0,03 con filtro sharpen. Esto caracteriza una falla por omisión: el modelo se vuelve excesivamente conservador y prácticamente no segmenta la lesión. Esta conducta explica los outliers mas extremos del scatter.

En U-Net ocurre un fallo distinto bajo el mismo régimen: el recall se mantiene muy alto (por ejemplo 96,3–99,0 % en ruido gaussiano con varianza 0,01 con filtro sharpen y ruido gaussiano con varianza 0,03 con filtro sharpen), pero la precisión cae abruptamente, llegando a valores cercanos a 51 % en ruido gaussiano con varianza 0,01 con filtro sharpen y cerca de 19 % en ruido gaussiano con varianza 0,03 con filtro sharpen. Esto indica sobre segmentación extrema, donde el modelo cubre casi toda la lesión (alto recall) pero incluye grandes áreas de tejido sano (baja precisión), provocando IoU y Dice muy bajos por exceso de máscara.

Mask R-CNN muestra mayor resistencia relativa, pero también sufre degradación marcada en el mismo régimen: en ruido gaussian con varianza 0,01 con filtro sharpen mantiene recall alrededor de 78–79 % y precisión alrededor de 70–71 %, mientras que en ruido gaussiano con varianza 0,03 con filtro sharpen cae a recall cercano a 48 % y precisión cercana a 40–41 %. En otras palabras, evita los extremos observados en DeepLabV3+ (omisión total) y U-Net (sobre segmentación extrema), pero igualmente queda bajo umbrales en condiciones críticas.

Un resultado consistente es que el suavizado (blur) tiende a estabilizar el comportamiento frente a ruido Gaussiano, mientras que el realce (sharpen) lo agrava. Por ejemplo, en condiciones como ruido gaussiano con varianza 0,01 con filtro blur y ruido gaussiano con varianza 0,03 con filtro blur las métricas se mantienen altas en los tres modelos (Precision en torno a 86–89 % y Recall típicamente sobre 83 % en DeepLabV3+ y sobre 90 % en U-Net y Mask R-CNN). En cambio, el esquema sharpen es el que concentra los mínimos de recall y precisión. Esto sugiere que el realce amplifica el ruido y distorsiona gradientes y texturas que los modelos utilizan para delimitar el borde de la lesión.

En síntesis, la combinación triple define el peor caso del sistema y evidencia que el riesgo

no depende solo de la resolución, sino de la interacción Ruido + Filtro. El caso mas adverso es ruido Gaussiano con sharpen, donde DeepLabV3+ falla por omisión (recall casi nulo), U-Net falla por sobre segmentación (precisión muy baja) y Mask R-CNN se degrada de manera importante. Por ello, para un despliegue confiable se recomienda evitar realce agresivo cuando exista ruido, o incorporar una etapa previa de control de calidad y mitigación de ruido antes de segmentar.

4.3.6. Evaluación de Generalización de la Robustez (Dataset PH2)

Para evaluar la capacidad de los modelos de transferir su robustez a un dominio externo (PH2), se aplicó el mismo sistema de degradación sistemática utilizado en ISIC 2018 y se analizó el rendimiento en cada condición.

Tabla 4.2: Comparación de métricas por modelo sobre PH2

Modelo	**IoU promedio**	**Dice promedio**	**Precision promedio**	**Recall promedio**
U-Net	0.8829	0.9354	0.9192	0.9608
DeepLabv3+	0.8511	0.9157	0.9506	0.896
Mask R-CNN	0.8299	0.9034	0.8474	0.9806

Previo a analizar la robustez ante degradaciones, se estableció un baseline de desempeño en el conjunto de prueba original de ISIC 2018 (sin alteraciones). La Tabla 4.1 resume las métricas promedio globales por modelo utilizando Intersection over Union (IoU), Dice Similarity Coefficient (Dice), Precision y Recall. Este baseline permite comparar posteriormente el impacto relativo de cada degradación, distinguiendo entre pérdida de solapamiento (IoU y Dice) y cambios en el tipo de error (precisión y recall).

En términos generales, U-Net presenta el comportamiento mas equilibrado entre solapamiento y estabilidad, con los mayores valores de IoU y Dice. DeepLabV3+ muestra un desempeño ligeramente inferior en solapamiento, acompañado de un recall menor, lo que sugiere una tendencia mas conservadora (mayor riesgo de omisión parcial). En contraste, Mask R-CNN alcanza un solapamiento similar a DeepLabV3+ pero con recall elevado y precision menor, lo que indica una estrategia de segmentación mas inclusiva (mayor probabilidad de sobre segmentación). Estas diferencias de perfil son relevantes, ya que bajo degradaciones severas pueden producir modos de falla distintos,

aspecto que se analiza en las secciones siguientes mediante los gráficos IoU vs Dice y recall/precision por condición.

4.3.6.1. Generalización Inicial (PH2 Sin Degradación)

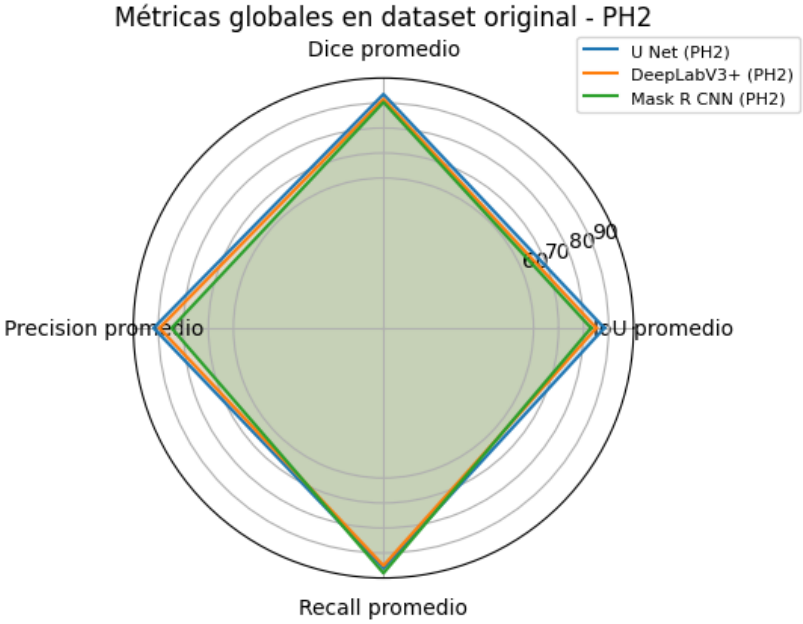


Fig. 4.24: Robustez Base (Gráfico Radar) PH2.

Para el dataset PH2 en su versión original (sin degradaciones), se incluye la figura 4.24 como resumen global del desempeño promedio por modelo. El gráfico radial permite comparar simultáneamente Intersection over Union (IoU), Dice Similarity Coefficient (Dice), precisión y recall, facilitando la identificación del perfil de error de cada arquitectura.

De manera general, los tres modelos se mantienen en un rango alto y relativamente cercano entre si, lo que sugiere que PH2 no introduce un cambio radical de dominio en condiciones ideales. Sin embargo, se observan diferencias de perfil: U-Net tiende a presentar un comportamiento mas equilibrado entre solapamiento (IoU y Dice) y métricas de tipo de error (precisión y recall). Mask R-CNN se caracteriza por un Recall mas elevado, lo que indica mayor cobertura de la lesión, pero con una precisión relativamente menor, consistente con una tendencia a sobre segmentación. DeepLabV3+ se mantiene cercano en solapamiento, con un balance intermedio entre precisión y recall. Este baseline es relevante porque anticipa que, bajo degradaciones, los modelos pueden

fallar de forma distinta: por omisión (caída de recall) o por exceso de máscara (caída de precisión), lo cual se analiza en las secciones siguientes.

4.3.6.2. Generalización ante Degradaciones Simples (PH2)

Los modelos se sometieron a los generadores de Reduccion de Resolución, ruido aditivo y Filtros aplicados.

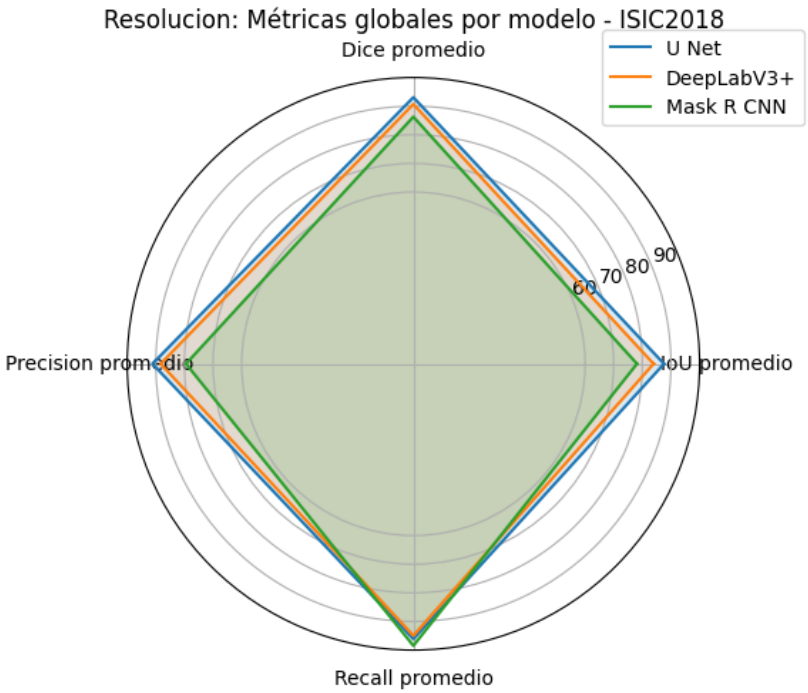


Fig. 4.25: Robustez ante resolución (Radar PH2).

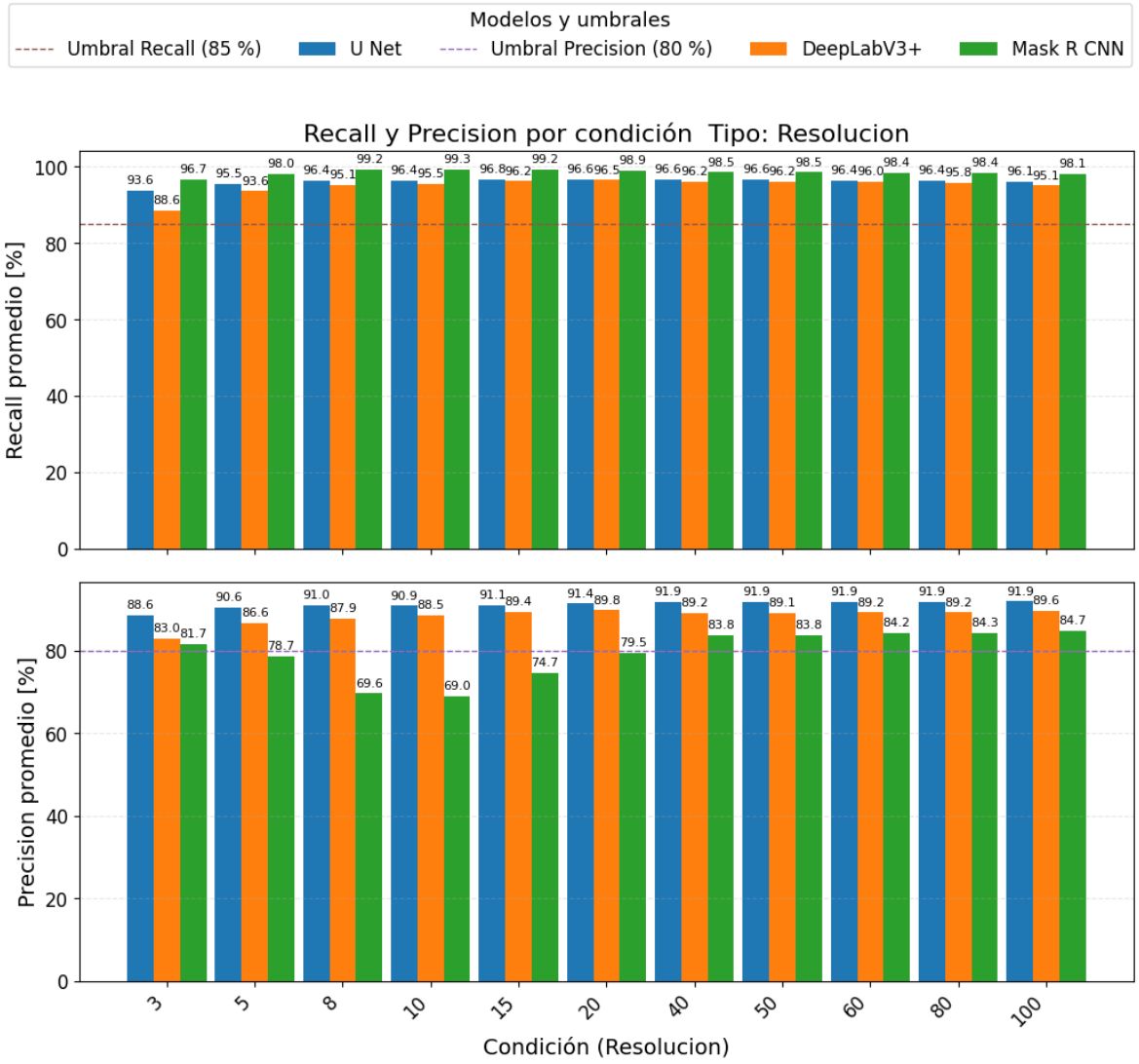


Fig. 4.26: Recall y precisión (Barras PH2).

Resistencia a la baja resolución La figura 4.25 resume el efecto global de la reduccion de resolucion sobre las metricas promedio. En terminos generales, el comportamiento se mantiene alto para los tres modelos, pero el perfil cambia segun la arquitectura: U-Net (U-shaped Network) conserva un balance estable entre Intersection over Union (IoU), Dice Similarity Coefficient (Dice), precisión y recall. DeepLabV3+ presenta un comportamiento cercano, aunque con valores levemente inferiores en cobertura en las resoluciones mas bajas. En cambio, Mask Region-based Convolutional Neural Network (Mask R-CNN) mantiene una cobertura elevada (recall alto), pero con mayor perdida relativa de precisión, consistente con sobre segmentacion cuando se

pierde detalle espacial.

Esta observacion se confirma en la Figura 4.26. En recall, los tres modelos se mantienen por sobre 88 % incluso en condiciones exigentes, y Mask R-CNN destaca por valores consistentemente muy altos (por ejemplo, 96.7 % en 3 % y sobre 98 % en la mayor parte del rango). Sin embargo, el efecto dominante aparece en precisión: Mask R-CNN cae de manera marcada en resoluciones bajas, alcanzando aproximadamente 69.6 % en 8 % y 69.0 % en 10 %, recuperandose progresivamente hacia resoluciones medias y altas (cercano a 83 %-85 % desde 40 % en adelante). En contraste, U-Net mantiene precisión alta y estable (aproximadamente 88.6 % a 91.9 %), y DeepLabV3+ se sostiene alrededor de 83.0 % a 89.6 %. En terminos operacionales, esta brecha indica que la reduccion extrema de resolucion afecta principalmente el exceso de mascara en Mask R-CNN, mas que la omision de lesion.

En comparacion con ISIC 2018 (International Skin Imaging Collaboration 2018), el patron es consistente (Mask R-CNN privilegia cobertura y penaliza precisión), pero en PH2 la caida de precisión en bajas resoluciones es mas pronunciada, lo que sugiere mayor sensibilidad a perdida de detalle fino bajo cambio de dominio.

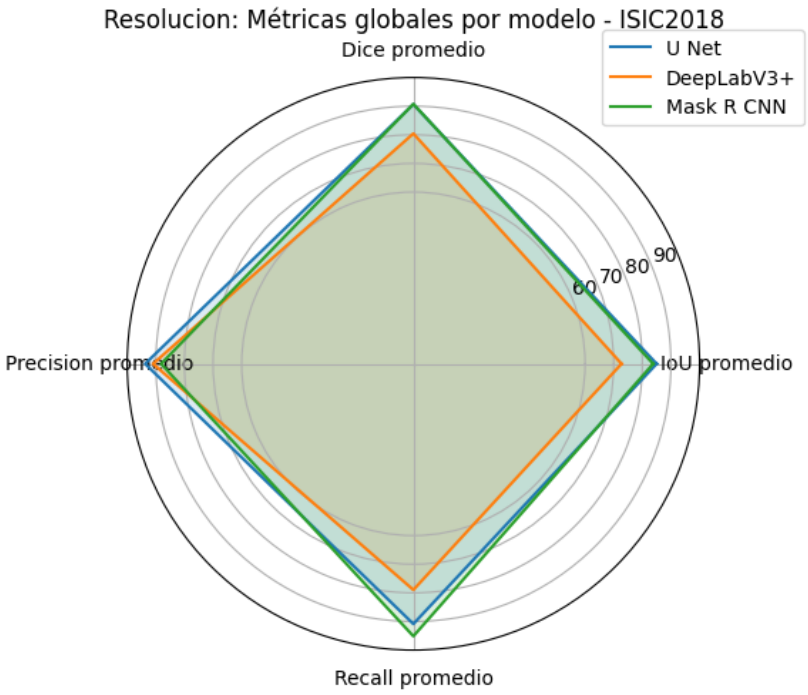


Fig. 4.27: Robustez ante ruido (Radar PH2).

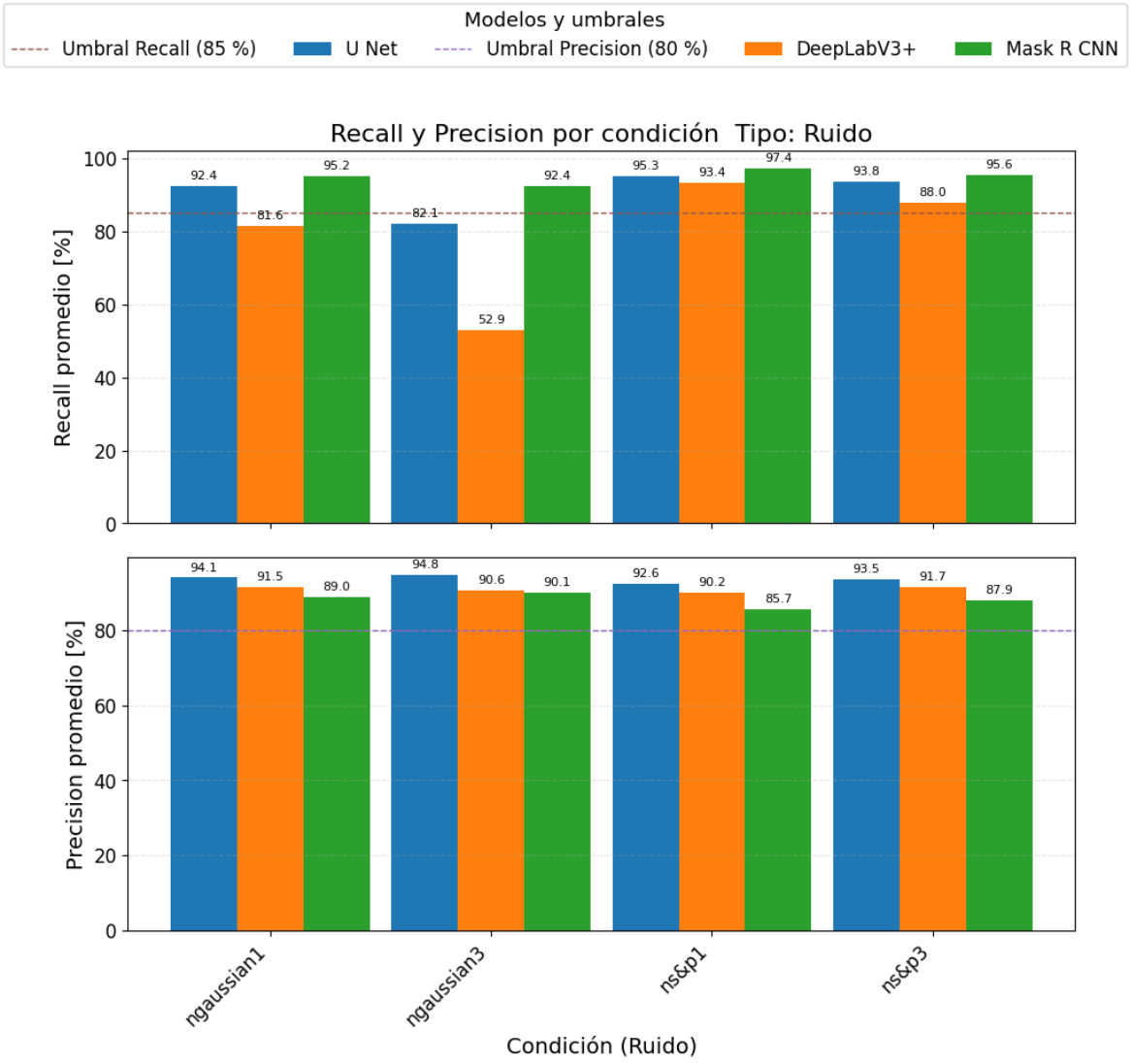


Fig. 4.28: Recall y precisión ante ruido (Barras PH2).

Vulnerabilidad ante ruido aditivo La figura 4.27 muestra que el ruido es la degradacion mas critica en PH2, ya que induce cambios fuertes en la cobertura (recall) para ciertas arquitecturas. El radar evidencia que el deterioro global no se distribuye de forma uniforme entre modelos: mientras Mask R-CNN mantiene cobertura alta, DeepLabV3+ reduce significativamente su desempeño promedio debido a escenarios de ruido severo.

En la figura 4.28 se observa el caso mas adverso en ruido gaussiano varianza 0.03, donde DeepLabV3+ cae a un recall cercano a 52.9 %, indicando sub segmentacion severa (omision de gran parte de la lesion). En ese mismo escenario, U-Net cae a

aproximadamente 82.1 %, quedando bajo un umbral operacional de 85 % si se aplica dicho criterio. En contraste, Mask R-CNN mantiene recall alto (alrededor de 92.4 %), mostrando mayor resiliencia en cobertura ante ruido gaussiano intenso.

Un punto relevante es que, incluso cuando cae el recall, la precisión permanece alta en DeepLabV3+ (por ejemplo, alrededor de 90.6 % en ruido gaussiano varianza 0.03). Esto indica que el modo de falla dominante bajo ruido severo es por omision (modelo mas conservador), mas que por sobre segmentacion. Para ruido tipo sal y pimienta, el comportamiento es mas estable: en ruido sal y pimienta los modelos mantienen recall alto (por ejemplo, U-Net 95.3 %, DeepLabV3+ 93.4 %, Mask R-CNN 97.4 %) con precision sobre 85 % en la mayoria de casos.

Al comparar con ISIC 2018, se mantiene la fragilidad de DeepLabV3+ ante ruido gaussiano intenso. Sin embargo, en ISIC 2018 U-Net no presentaba una caida comparable en ruido gaussiano varianza 0.03, mientras que en PH2 si cae a 82.1 %. Esto sugiere que el cambio de dominio hace mas exigente el escenario de ruido severo, reduciendo el margen operacional incluso para modelos que fueron mas estables en ISIC 2018.

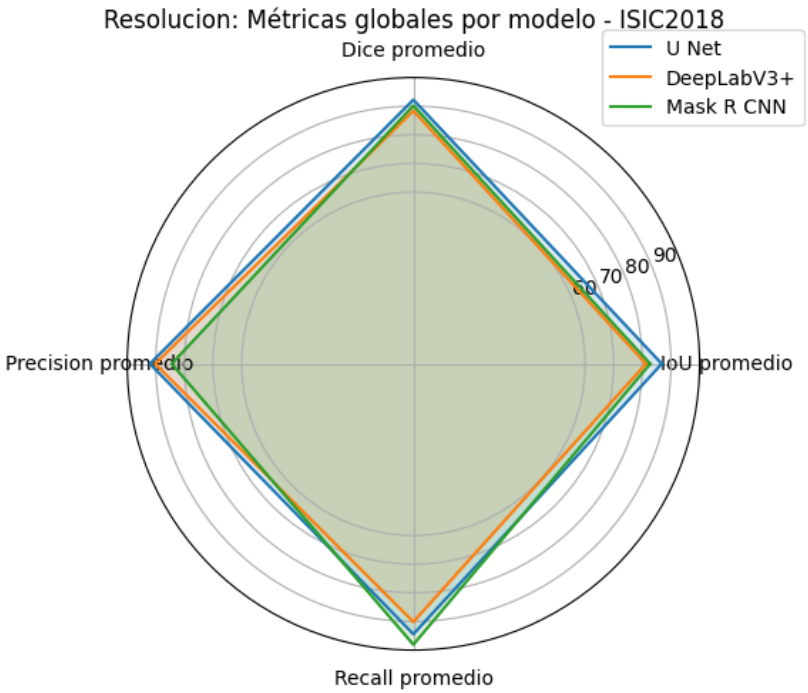


Fig. 4.29: robustez ante filtros (Radar PH2).

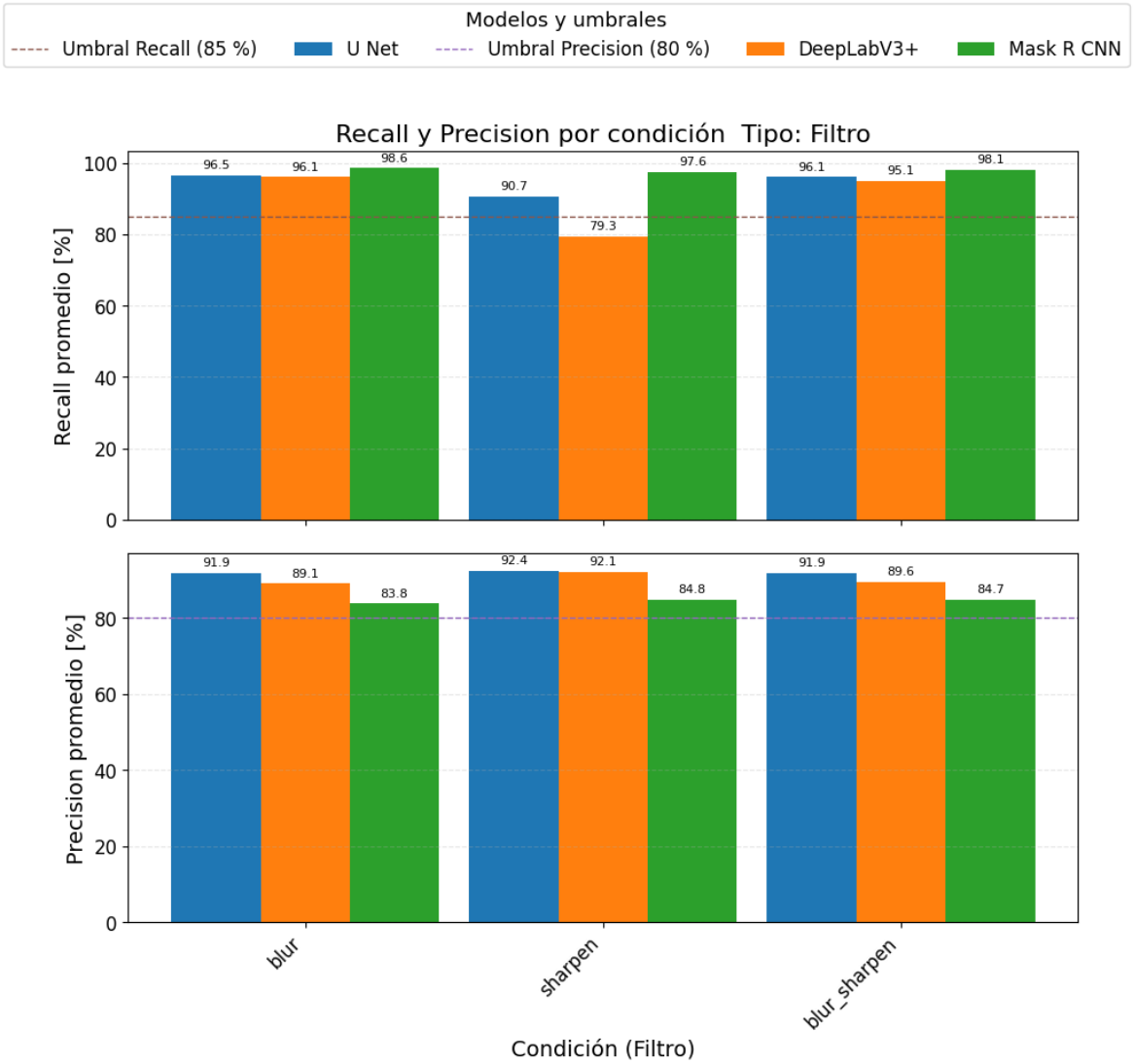


Fig. 4.30: recall y precisión ante Filtros (Barras PH2).

Efecto de los filtros La figura 4.29 evidencia que los filtros afectan de manera asimetrica segun la operacion aplicada. Globalmente, el suavizado (blur) no destruye la capacidad de localizar la lesion, mientras que el realce (sharpen) puede introducir artefactos que degradan la segmentacion, especialmente en DeepLabV3+.

Esta tendencia se ve con claridad en la figura 4.30. En blur y blur_sharpen, los tres modelos mantienen recall alto (por ejemplo, en blur: U-Net 96.5 %, DeepLabV3+ 96.1 %, Mask R-CNN 98.6 %). En cambio, en sharpen DeepLabV3+ cae a un recall

cercano a 79.3 %, por debajo de un umbral operacional de 85 %, lo que refleja sub segmentacion bajo realce de bordes. U-Net y Mask R-CNN se mantienen robustos en esa condicion (U-Net alrededor de 90.7 %, Mask R-CNN alrededor de 97.6 %).

En precision, U-Net y DeepLabV3+ se mantienen altos (aproximadamente 89 % a 92 %), mientras Mask R-CNN se mantiene mas bajo (cercano a 83.8 % a 84.8 %), consistente con su tendencia a sobre segmentar de forma relativa. En comparacion con ISIC2018, sharpen tambien era la condicion mas desfavorable para DeepLabV3+, pero en PH2 el deterioro es aun mas evidente (recall cercano a 79.3 %), lo que refuerza que aplicar realce como preprocesamiento no es recomendable si se busca robustez inter dataset.

En sintesis, PH2 muestra patrones consistentes con ISIC2018, pero con degradaciones mas severas en condiciones especificas: (i) ruido gaussiano intenso, donde DeepLabV3+ y tambien U-Net pueden caer bajo el umbral de recall, y (ii) resoluciones extremadamente bajas, donde Mask R-CNN mantiene alta cobertura pero pierde precision de forma marcada. Por esta razon, el umbral operacional propuesto resulta util para separar condiciones aceptables de condiciones limite, distinguiendo entre fallas por omision (caida de recall) y fallas por sobre segmentacion (caida de precision).

4.3.6.3. Generalización ante Degradaciones Combinadas (PH2)

En esta sección se analizan degradaciones dobles aplicadas al conjunto PH2, utilizando como criterio de operación los umbrales definidos para: IoU (Intersection over Union) $\geq$ 75 %, Dice (Dice Similarity Coefficient) $\geq$ 85 %, recall (sensibilidad) $\geq$ 85 % y precisión $\geq$ 80 %.

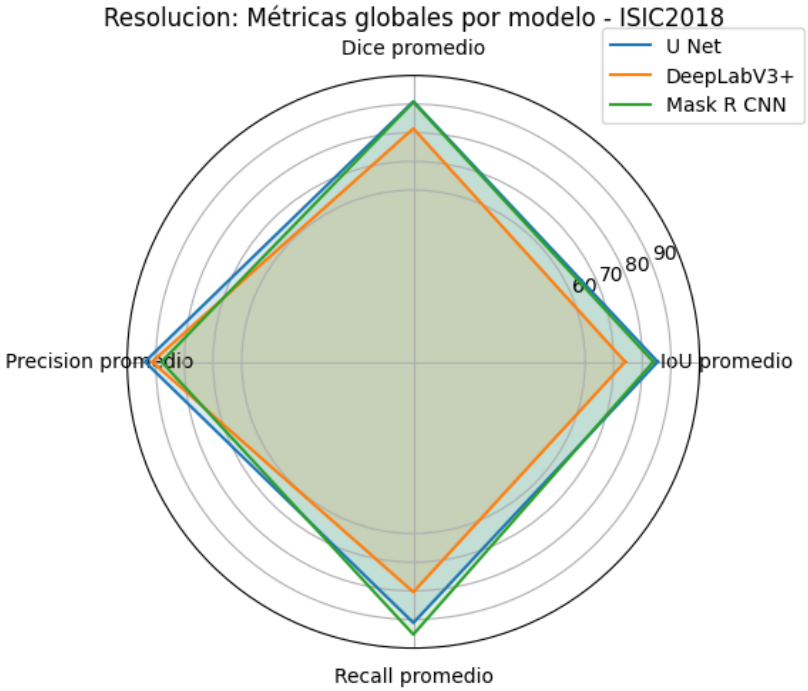


Fig. 4.31: Robustez ante doble combinación resolución + ruido (Radar PH2).

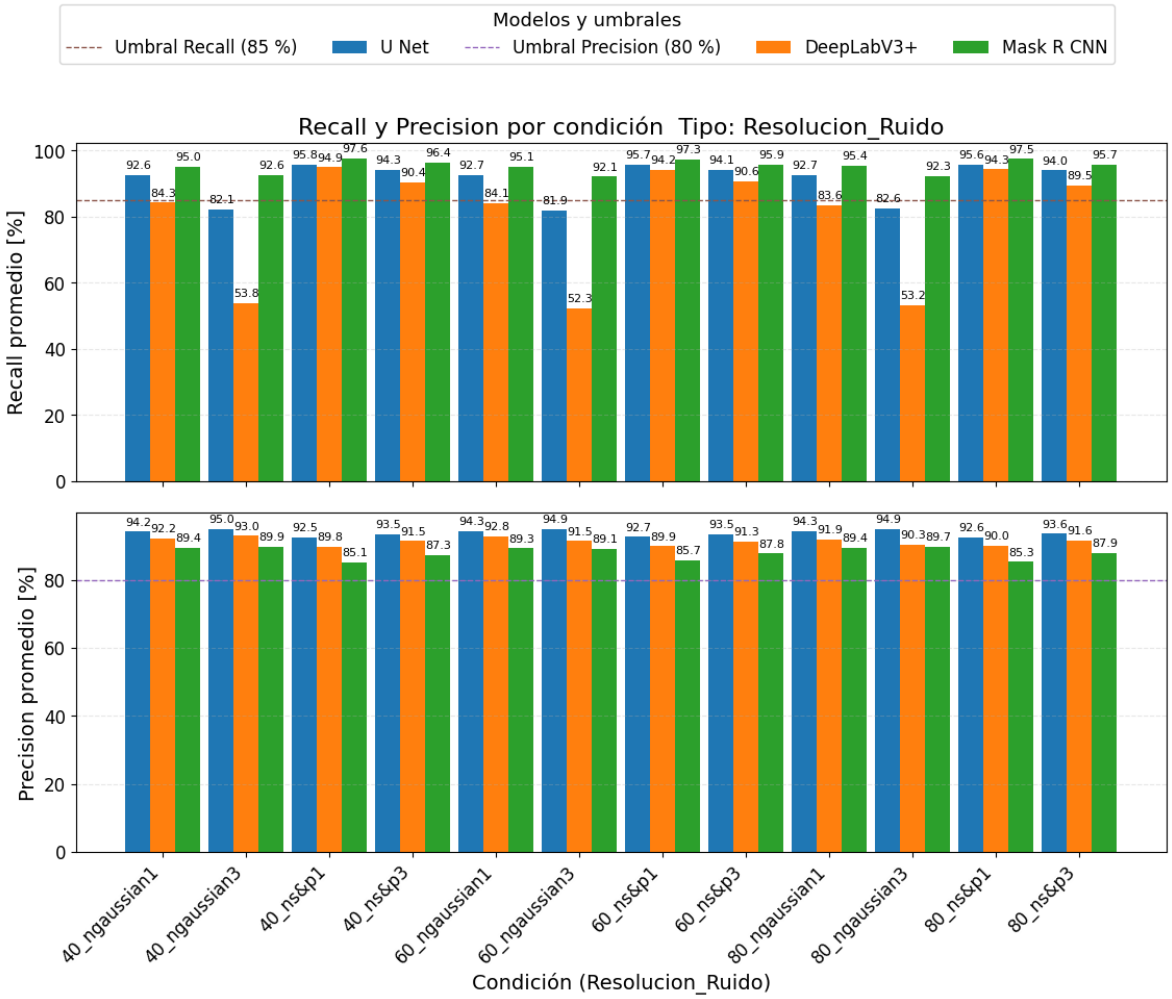


Fig. 4.32: Recall y precisión ante doble combinación resolución + ruido (Barras PH2).

Combinación doble (resolución + ruido) En la combinación resolución + ruido, el comportamiento está dominado principalmente por el tipo e intensidad de ruido, mientras que la resolución actúa como un factor secundario. Bajo condiciones de ruido gaussiano severo (por ejemplo, ruido gaussiano con varianza de 0.03), se observa una caída marcada en recall (sensibilidad), especialmente en DeepLabV3+, llegando a valores bajo el umbral operacional en múltiples condiciones. U-Net también muestra disminuciones de recall bajo ruido severo, aunque en menor medida. En contraste, Mask R-CNN (Region-based Convolutional Neural Network) tiende a mantener recall (Sensibilidad) alto y estable, superando el umbral en la mayoría de las combinaciones.

Por otro lado, la precisión se mantiene elevada en gran parte de las condiciones, lo cual

puede generar una interpretación engañosa si se analiza de forma aislada. El patrón de precisión alta junto a recall bajo es consistente con un fenómeno de subsegmentación: el modelo predice menos píxeles de lesión, reduciendo falsos positivos, pero incrementando falsos negativos. En segmentación de lesiones, este comportamiento es crítico, ya que omitir parte de la lesión reduce la utilidad clínica del resultado. Por ello, el criterio operacional permite marcar explícitamente qué combinaciones dejan de ser confiables.

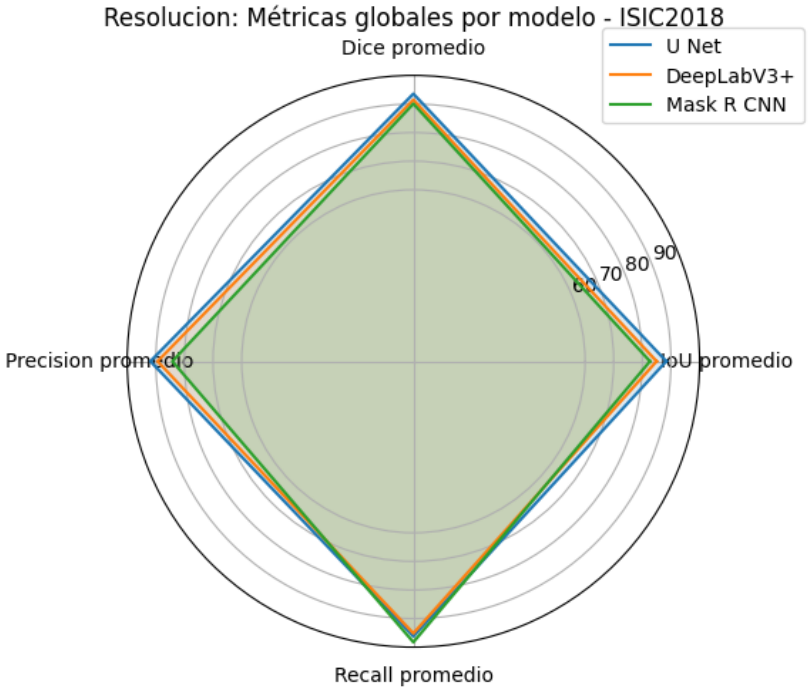


Fig. 4.33: Robustez ante doble combinación resolución + filtro (Radar PH2).

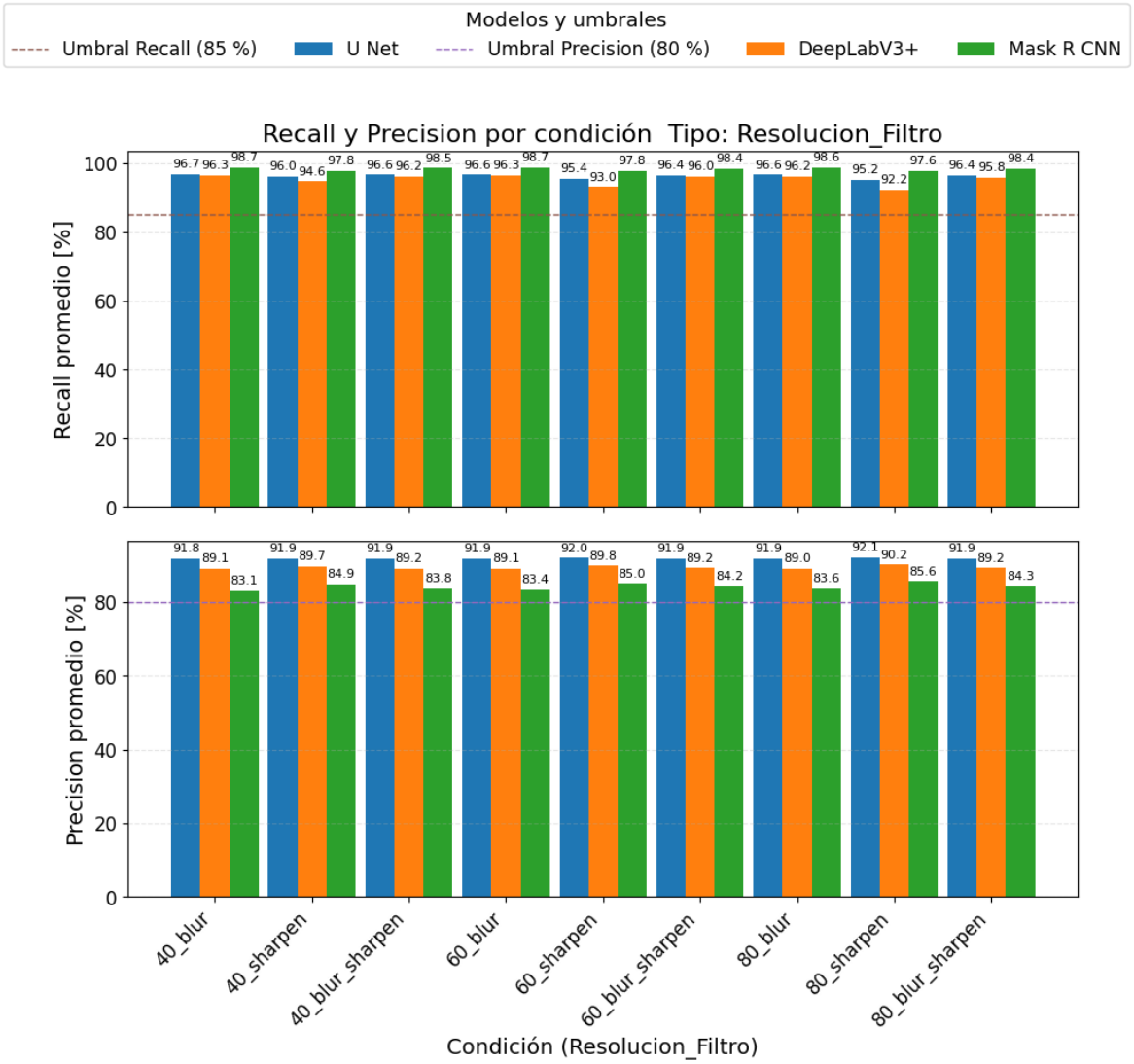


Fig. 4.34: Recall y precisión ante doble combinación resolución + filtro (Barras PH2).

Combinación doble (resolución + filtro) En **resolución + filtro**, el desempeño se mantiene globalmente estable en comparación con combinaciones que incluyen ruido. En general, los modelos conservan recall (sensibilidad) por sobre el umbral operacional a lo largo de blur, sharpen y blur_sharpen, incluso variando la resolución. Esto sugiere que, en ausencia de ruido severo, los modelos presentan robustez frente a cambios moderados de nitidez y suavizado, y que la degradación por resolución dentro del rango evaluado no introduce fallas graves de cobertura de lesión.

Las diferencias se observan principalmente en precisión). En particular, Mask R-CNN (Region-based Convolutional Neural Network) tiende a quedar más cercano al

umbral de Precision en diversas condiciones, lo cual sugiere una mayor propensión a sobresegmentación (incremento de falsos positivos) en comparación con U-Net y DeepLabV3+. Aun así, al permanecer sobre el umbral, estas combinaciones se consideran operables bajo el criterio definido.

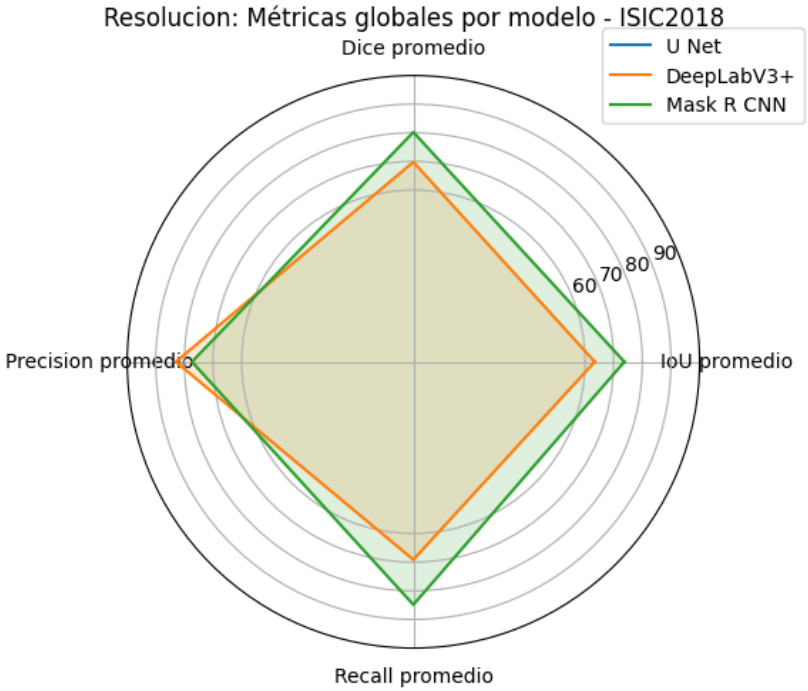


Fig. 4.35: Robustez ante doble combinación ruido + filtro (Radar PH2).

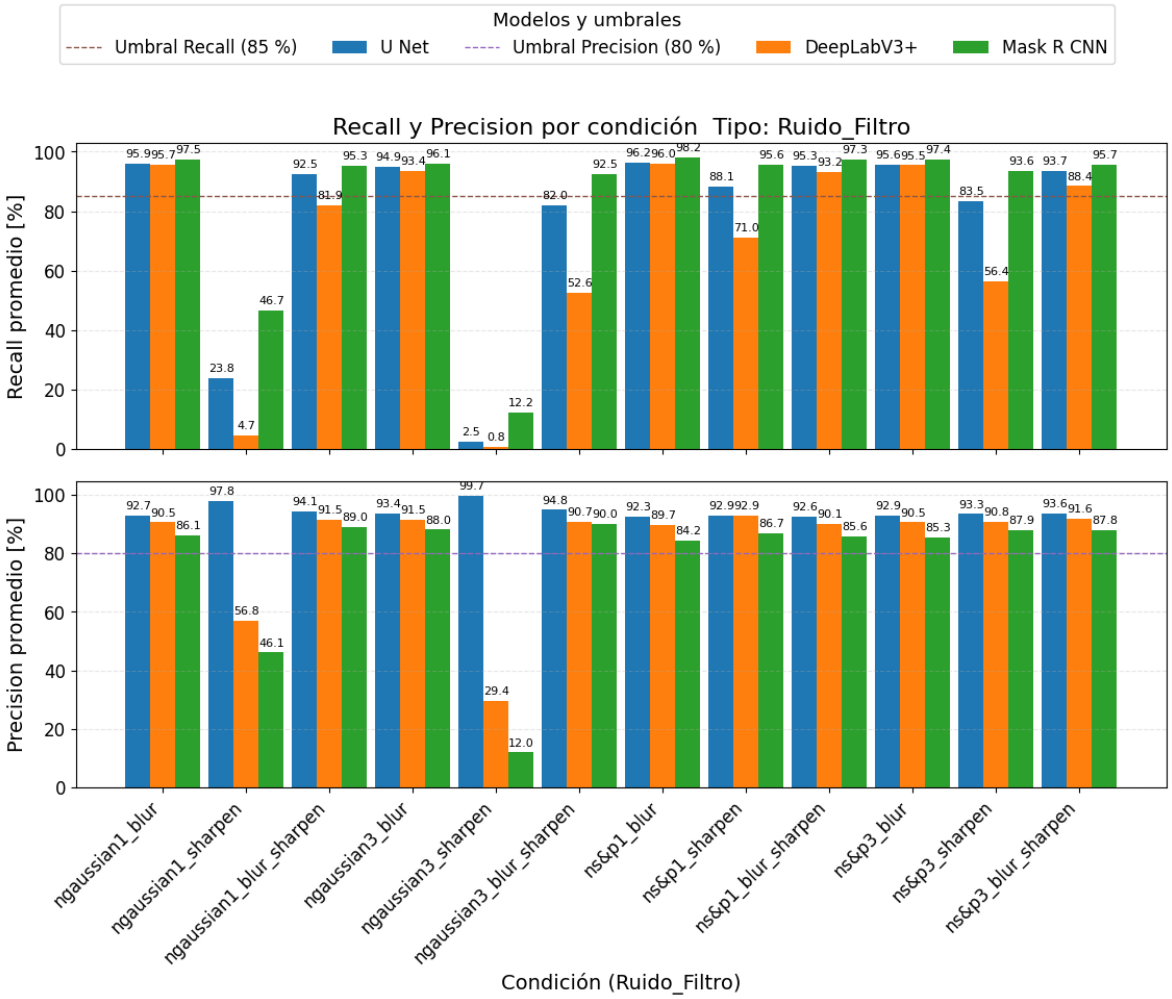


Fig. 4.36: Recall y precisión ante doble combinación ruido + filtro (Barras PH2).

Combinación doble (ruido + filtro) La combinación ruido + filtro es la más exigente, debido a efectos de interacción entre ambos tipos de degradación. En particular, aplicar sharpen bajo ruido gaussiano (por ejemplo, ruido gaussiano con varianza 0.01 y ruido gaussiano con varianza 0.03) provoca deterioros severos: el recall (sensibilidad) puede colapsar a valores muy inferiores al umbral, con un impacto especialmente fuerte en DeepLabV3+. Además, es posible observar precision relativamente alta incluso cuando recall es muy bajo, lo cual es consistente con el caso extremo de subsegmentación, donde el modelo predice muy pocas regiones positivas (pocos falsos positivos, pero muchos falsos negativos).

En cambio, cuando se aplica blur (o blur_sharpen) el desempeño tiende a recuperarse,

lo cual es coherente con el efecto de suavizado de blur al reducir textura espuria antes del realce. Bajo este resultado, el criterio operacional respalda una regla práctica: evitar sharpen cuando existe ruido, y preferir estrategias de suavizado si el escenario es ruidoso. Esta combinación, además, valida la utilidad del umbral operacional como detector de zonas no confiables, ya que las fallas son suficientemente grandes como para no dejar ambigüedad.

Síntesis de operabilidad Bajo el criterio operacional (IoU (Intersection over Union) $\geq 75\,\%$, Dice (Dice Similarity Coefficient) $\geq 85\,\%$, recall (sensibilidad) $\geq 85\,\%$, precisión $\geq 80\,\%$), las combinaciones dobles muestran que: (i) **Resolución + filtro** se mantiene mayoritariamente operable, (ii) **Resolución + ruido** falla principalmente por caídas de recall y sensibilidad bajo ruido gaussiano severo, (iii) **Ruido + filtro** presenta las fallas más críticas por interacción negativa, especialmente cuando sharpen amplifica el efecto del ruido. En consecuencia, el umbral operacional permite delimitar objetivamente el dominio donde las predicciones pueden considerarse confiables y comparables entre modelos.

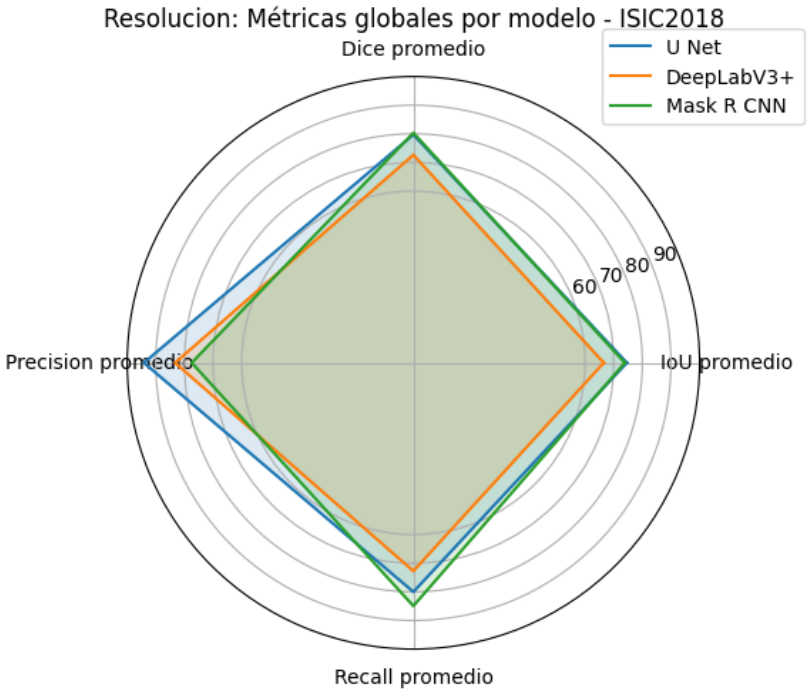


Fig. 4.37: Robustez ante triple combinación (radar PH2).

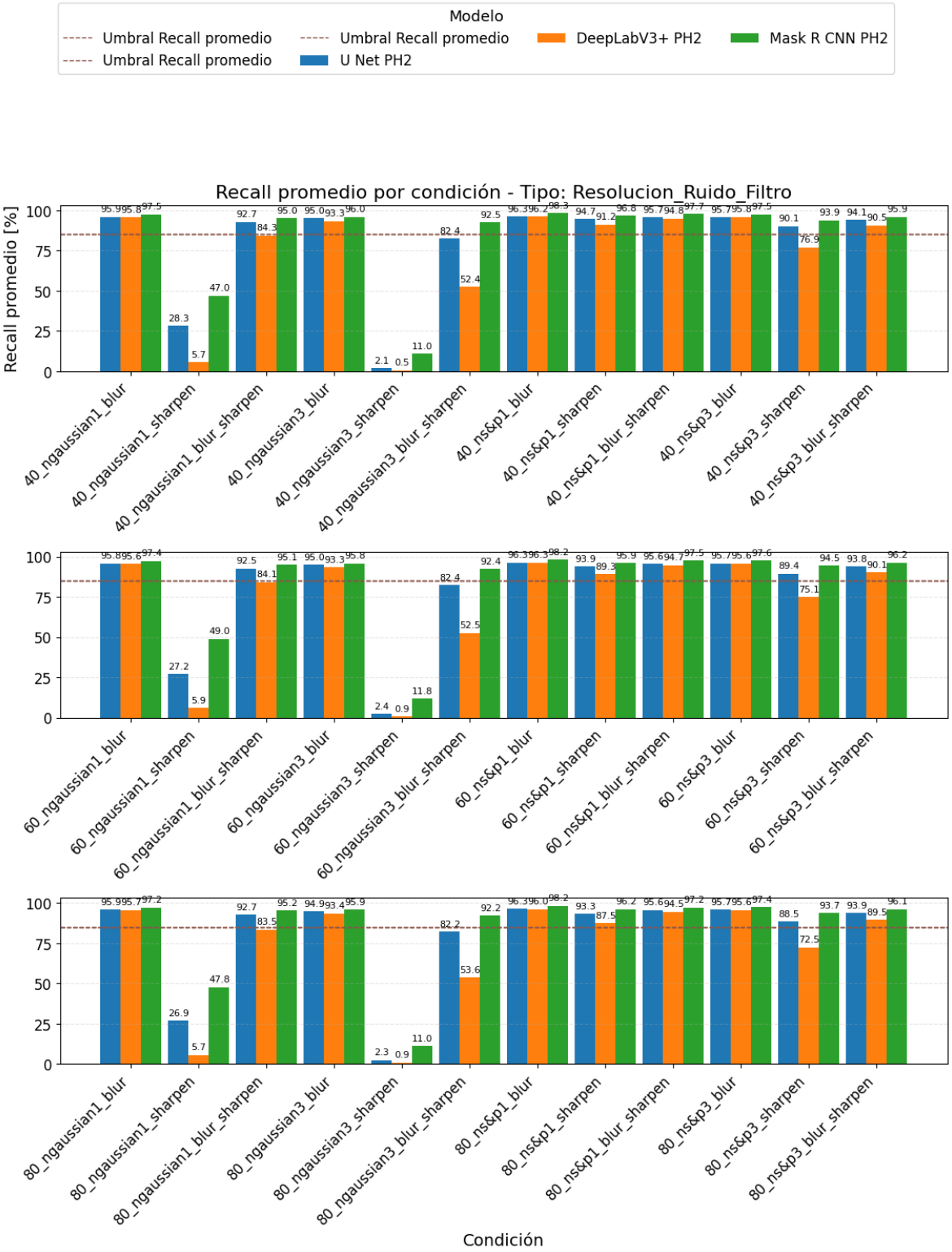


Fig. 4.38: Recall ante triple combinación (barras PH2).

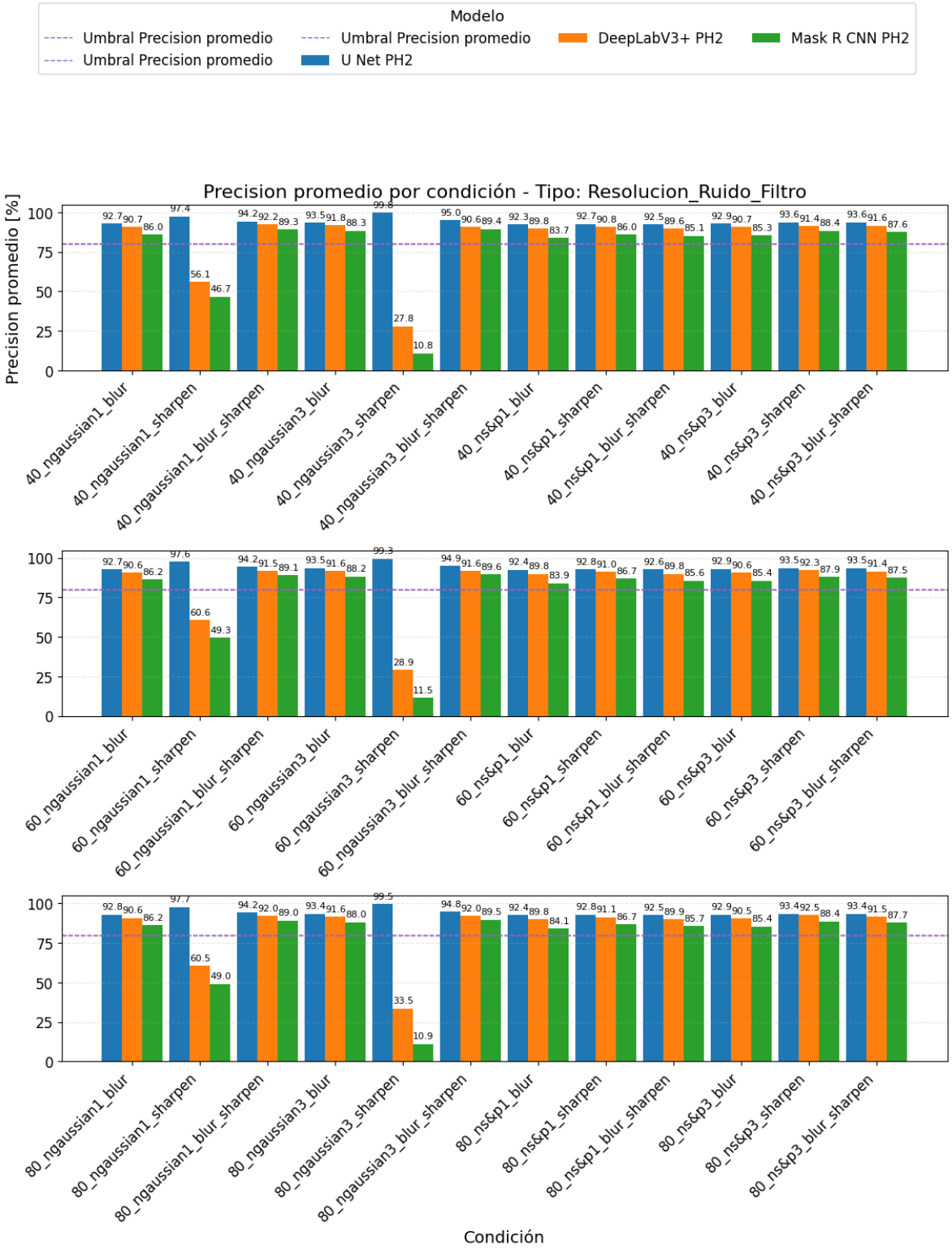


Fig. 4.39: Precisión ante triple combinación (barras PH2).

Combinación triple (resolución + ruido + filtro) La figura 4.37 evidencia que, al combinar simultáneamente resolución, ruido y filtro, el desempeño global cae de forma marcada respecto a degradaciones simples. En particular, IoU y Dice se desplazan hacia valores medios, indicando pérdida de solapamiento y contornos menos precisos. El radar también revela un compromiso claro entre precision y recall: U-Net concentra su fortaleza en precision (segmenta con menos falsos positivos), mientras que Mask R-CNN tiende a privilegiar recall (recupera mayor proporción de la lesión), y DeepLabV3+ se ubica por debajo en la mayoría de ejes, mostrando mayor sensibilidad a la degradación compuesta.

En la combinación triple, la precisión se mantiene relativamente alta cuando el modelo adopta una estrategia conservadora (segmenta menos área, pero con mayor certeza). Esto favorece a U-Net, que destaca en precision incluso cuando IoU y Dice ya disminuyen. Mask R-CNN exhibe precisión más moderada, coherente con su tendencia a recuperar más área (mayor recall), lo que naturalmente incrementa el riesgo de falsos positivos en condiciones altamente degradadas. DeepLabV3+ muestra mayor variabilidad, indicando que su frontera de decisión se vuelve menos estable frente a ruido y perturbaciones del borde.

El recall es el indicador que más sufre en escenarios triple cuando la señal visual se vuelve ambigua (ruido) y el borde pierde definición (filtro) bajo menor resolución. En este contexto, Mask R-CNN suele conservar mejor el recall, pero con la penalización mencionada en precisión. U-Net reduce recall al priorizar precisión, lo que se interpreta como subsegmentación parcial en casos difíciles. DeepLabV3+ presenta las mayores caídas, reforzando que es el modelo menos robusto a degradaciones compuestas en PH2.

En ISIC 2018, las tendencias generales suelen ser más favorables por la escala del conjunto y la diversidad de entrenamiento, reflejándose en métricas globales más altas y una degradación más gradual. En PH2, el cambio de dominio (menos datos, diferencias de adquisición y apariencia) hace que la combinación triple empuje más rápido a los modelos fuera de su zona estable. Aun así, el patrón relativo se mantiene: U-Net destaca en precisión, Mask R-CNN en recall y DeepLabV3+ es el más sensible a degradaciones fuertes.

Dado que la combinación triple concentra el peor caso, es esperable que varios escenarios

no cumplan simultáneamente umbrales estrictos (IoU, Dice, precision y recall). Esto no invalida el umbral, sino que lo convierte en un criterio de seguridad: si el caso cae bajo umbral, el sistema debe reportar baja confianza (o requerir revisión), evitando decisiones automáticas sobre imágenes cuya calidad efectiva ya está fuera del rango donde los modelos fueron confiables.

4.4. Discusión y conclusiones de robustez operacional

El análisis exhaustivo de los resultados en los **58** escenarios de degradación sistemática y la validación en el dominio externo PH2 permiten extraer conclusiones sólidas sobre la robustez y los límites operativos de las arquitecturas implementadas.

1. **Modelo más robusto (Mask R-CNN):** Mask R-CNN demostró ser la arquitectura más robusta en el rendimiento global, manteniendo el mayor recall y la mayor estabilidad en escenarios combinados. Su enfoque basado en segmentación de instancias le permite aislar la lesión de manera más efectiva, siendo menos susceptible a los artefactos distribuidos en el fondo de la imagen, superando consistentemente a las arquitecturas de segmentación semántica (U-Net y DeepLabv3+).

2. **Vulnerabilidad crítica (ruido gaussiano):** La arquitectura DeepLabv3+ presentó el punto de fallo más temprano y crítico, particularmente ante el **ruido gaussiano nivel 3** (0.03). Esta condición provocó una caída de la sensibilidad (recall) a **51**,**6** % (ver Figura 4.12 y 4.28), indicando una alta tasa de Falsos Negativos, lo cual es inaceptable en un contexto clínico.

3. **Condición límite (fallo catastrófico):** La combinación de ruido gaussiano nivel 3 y el filtro sharpen se identificó como la condición letal que provoca el fallo catastrófico en los tres modelos en ambos dominios (ISIC y PH2). Esto establece un límite físico para la calidad mínima de imagen.

4. **Umbral de resolución:** Los tres modelos son extremadamente tolerantes a la baja resolución. La segmentación se mantuvo confiable (sobre el umbral clínico) hasta el nivel más bajo probado (resolución **3** % del tamaño original), lo cual es

validado por los resultados de PH2.

5. **Capacidad de generalización de la robustez (dataset PH2):** La evaluación exhaustiva sobre el dataset PH2 (Sección 4.3.6) confirmó la transferibilidad de los hallazgos de robustez. La superioridad de Mask R-CNN y la vulnerabilidad crítica de DeepLabv3+ al ruido gaussiano se mantuvieron en este dominio externo, asegurando que las conclusiones de robustez operacional son aplicables a diferentes fuentes de imágenes dermatoscópicas.

En resumen, si bien los tres modelos demuestran una excelente capacidad de segmentación en imágenes limpias, solo Mask R-CNN demostró la robustez operacional necesaria para un entorno de telemedicina donde la calidad de la imagen no puede ser controlada estrictamente.

Capítulo 5

Conclusiones

5.1. Síntesis del trabajo

En este trabajo se desarrolló y evaluó un marco experimental para la segmentación automática de lesiones cutáneas mediante Aprendizaje Profundo (Deep Learning), comparando tres arquitecturas representativas: U-Net, DeepLabV3+ y Mask Region-based Convolutional Neural Network (Mask R-CNN). La evaluación se realizó principalmente sobre el conjunto de prueba del dataset ISIC 2018, incorporando además el dataset PH2 como validación adicional de generalización.

El aporte central del estudio es el análisis de robustez ante degradaciones controladas de calidad de imagen, considerando reducción de resolución, incorporación de ruido y aplicación de filtros, además de combinaciones dobles y una combinación triple. Para caracterizar el desempeño se utilizaron métricas de solapamiento (Intersection over Union, IoU, y Dice Similarity Coefficient, Dice), junto con métricas de tipo de error (Precision y Recall), permitiendo distinguir entre omisión de la lesión y sobre segmentación.

5.2. Conclusiones principales

A partir de los resultados presentados en el capítulo anterior, se concluye lo siguiente:

1. **Desempeño base y perfil de error.** En imágenes sin degradación del dataset ISIC 2018, U-Net obtiene el desempeño mas equilibrado (IoU = 0.8284, Dice = 0.893, Precision = 0.9156, Recall = 0.9044), lo que indica segmentaciones consistentes y con bajo sesgo hacia omisión o exceso. DeepLabV3+ presenta menor desempeño promedio (IoU = 0.7867, Dice = 0.8631) con un perfil mas conservador, reflejado en Recall menor. Mask R-CNN alcanza IoU y Dice comparables a DeepLabV3+ (IoU = 0.788, Dice = 0.8688), pero con un perfil característico de alta cobertura (recall = 0.9445) y menor precisión (0.8351), lo que evidencia una tendencia a sobre segmentación. Este resultado confirma que la selección del modelo depende del criterio operacional: minimizar omisiones (recall) o minimizar falsos positivos (precisión).

2. **La reducción de resolución, por si sola, no es la degradación mas crítica en el rango evaluado.** En degradación simple por resolución, los modelos mantienen un comportamiento estable hasta niveles muy bajos, y el quiebre se concentra en el caso extremo (por ejemplo 3 %). En ese escenario, DeepLabV3+ cruza claramente bajo los umbrales de solapamiento, y Mask R-CNN queda cerca del límite, lo que sugiere que la pérdida de detalle de borde afecta mas a arquitecturas sensibles al contorno. Aun así, en la mayoría de niveles el recall se mantiene alto, indicando que el problema dominante es la precisión del borde mas que la desaparición completa de la segmentación.

3. **El ruido, especialmente el ruido Gaussiano, es el factor mas determinante para la robustez.** En degradación simple por ruido se observan outliers y colapsos severos en el plano IoU vs Dice, lo que evidencia un cambio de régimen desde segmentación estable a fallas estructurales. DeepLabV3+ se degrada principalmente por omisión (caídas fuertes de recall bajo ruido Gaussiano), mientras que U-Net tiende a fallar por sobre segmentación (precisión baja cuando el ruido es severo). Mask R-CNN se mantiene mas estable en promedio, pero también presenta deterioro cuando el ruido se intensifica. En contraste, el ruido tipo sal y pimienta genera un impacto mas moderado y con mayor estabilidad relativa entre condiciones.

4. **Los filtros aislados tienden a ser una degradación secundaria, pero su interacción con ruido puede amplificar fallas.** En degradación simple por filtro, la mayoría de condiciones se mantiene sobre umbrales, indicando

que blur y sharpen no provocan por si solos un colapso generalizado. Sin embargo, al combinar ruido con filtro se observa el efecto mas crítico del estudio: la combinación de ruido Gaussiano con sharpen concentra los mínimos de desempeño, generando fallas catastróficas. En esta condición, DeepLabV3+ falla por omisión (Recall cercano a cero), U-Net falla por sobre segmentación extrema (precisión muy baja) y Mask R-CNN también se degrada de forma marcada. Este patrón confirma que el realce (sharpen) puede amplificar artefactos del ruido y distorsionar gradientes que los modelos usan para delimitar bordes.

5. **En degradaciones combinadas dobles y en la combinación triple, el comportamiento no es aditivo y aparecen puntos de quiebre.** En combinaciones dobles, Resolución + Filtro se comporta como un escenario relativamente estable dentro del rango evaluado, mientras que Resolución + Ruido y Ruido + Filtro heredan la inestabilidad del ruido, con quiebres pronunciados bajo ruido Gaussiano intenso. En la combinación triple (Resolución + Ruido + Filtro) se observan claramente dos regímenes: un conjunto de condiciones estables cercanas a la zona aceptable y un conjunto de outliers severos bajo ruido Gaussiano con sharpen, lo que representa el peor caso operacional del sistema.

5.3. Justificación y utilidad del umbral operacional

Para comparar robustez y determinar condiciones límite se definió un umbral operacional basado en métricas de solapamiento y de tipo de error: IoU $\geq$ 0,75, Dice $\geq$ 0,85, recall $\geq$ 0,85 y precisión $\geq$ 0,80. Este criterio es coherente matemáticamente, ya que IoU y Dice se relacionan por:

$$Dice = \frac{2 \cdot IoU}{1 + IoU}. \tag{5.1}$$

De esta relación se obtiene que $IoU = 0{,}75$ corresponde aproximadamente a $Dice \approx 0{,}857$, lo que justifica el valor $Dice \geq 0{,}85$ como criterio consistente y de interpretación simple. Complementariamente, recall y precisión permiten caracterizar el modo de falla: Recall controla omisiones (falsos negativos) y precisión controla sobre segmentación (falsos positivos). En los resultados, el umbral operacional separa de manera clara

el régimen estable del régimen de colapso, especialmente en combinaciones con ruido Gaussiano y sharpen, donde las caídas son abruptas y no marginales.

5.3.1. Generalización y robustez en PH2

La evaluación sobre PH2 permite estudiar la robustez de generalización del sistema ante un cambio de dominio. PH2 difiere de ISIC2018 en condiciones de adquisición, distribución de colores, contraste, características de la piel y composición del dataset, por lo que es esperable observar una brecha de generalización. Aun así, los resultados muestran que los modelos mantienen patrones de comportamiento comparables a ISIC 2018 en condiciones moderadas, lo que sugiere que las representaciones aprendidas capturan rasgos relevantes de lesiones que se transfieren entre datasets.

El análisis global en PH2 evidencia que la comparación entre arquitecturas conserva tendencias clave. U-Net mantiene una conducta robusta en términos de estabilidad, destacando por un equilibrio favorable entre solapamiento y errores en un rango amplio de condiciones. Mask R-CNN mantiene una ventaja relativa en recall, lo que se interpreta como mayor sensibilidad para incluir la lesión completa, aunque con un tradeoff hacia menor precisión en escenarios específicos. DeepLabV3+ muestra una sensibilidad mayor cuando el dominio cambia y las degradaciones se combinan: en condiciones críticas puede presentar caídas pronunciadas en recall, lo cual es consistente con un aumento de falsos negativos cuando el borde de la lesión pierde definición.

En degradaciones simples sobre PH2, la reducción de resolución y el filtrado tienden a degradar de forma gradual, mientras que el ruido es el factor que más presiona la generalización. En combinaciones dobles se refuerza el efecto amplificador entre ruido y realce, y en la combinación triple se observan los casos de mayor inestabilidad. Estas observaciones permiten concluir que PH2 confirma el rol de las degradaciones combinadas como condición adversa y valida el uso de umbrales como criterio independiente del dataset: el umbral no busca optimizar el rendimiento promedio, sino proteger el sistema cuando la calidad de entrada lo vuelve impredecible.

5.4. Implicancias prácticas

Los hallazgos permiten proponer recomendaciones operacionales para un despliegue confiable:

- **Control de calidad de entrada:** incorporar un filtro de calidad (quality gate) que estime nivel de ruido y evite operar en condiciones adversas, o que active una etapa de mitigación.
- **Mitigación de ruido antes de segmentar:** si existe ruido Gaussiano, se recomienda aplicar denoising o suavizado controlado. En particular, evitar realce agresivo (sharpen) cuando hay ruido, ya que amplifica fallas.
- **Selección del modelo según riesgo:** si la prioridad es no omitir la lesión, Mask R-CNN ofrece mayor recall, aunque puede sobre segmentar. Si se busca equilibrio global y estabilidad en múltiples condiciones, U-Net se posiciona como la opción mas consistente en este estudio. DeepLabV3+ requiere mayor cuidado ante ruido Gaussiano por su tendencia a omisión.

5.5. Limitaciones y trabajo futuro

Entre las principales limitaciones se encuentra que el umbral operacional es un criterio técnico y no clínico, y que las degradaciones evaluadas no cubren todos los artefactos reales posibles en dermatoscopia. Además, las conclusiones dependen del conjunto de datos y del protocolo experimental implementado.

Como líneas de trabajo futuro se proponen:

1. **Ampliar degradaciones realistas:** pelo, burbujas, viñeteo, desenfoque por movimiento, variaciones de iluminación y compresión.
2. **Preprocesamiento adaptativo:** seleccionar automáticamente la estrategia de denoising o normalización según el tipo y severidad de degradación detectada.
3. **Incertidumbre y confiabilidad:** incorporar estimación de incertidumbre para marcar predicciones poco confiables y mejorar la seguridad del sistema.

4. **Validación con expertos:** contrastar umbrales operacionales con criterios de aceptación clínica, especialmente en escenarios donde el área de la lesión impacta decisiones posteriores.
5. **Evaluación de costo computacional:** medir latencia, memoria y estabilidad en hardware objetivo para despliegue.

Bibliografía

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, «Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation», European Conference on Computer Vision (ECCV), Munich, Alemania, 2018.

[2] K. He, G. Gkioxari, P. Dollár, R. Girshick, «Mask R-CNN», IEEE International Conference on Computer Vision (ICCV), Venice, Italia, 2017.

[3] N. Codella, V. Rotemberg, P. Tschandl, et al., «Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)», arXiv:1902.03368, Febrero 2019.

[4] L. Bi, J. Kim, E. Ahn, A. Kumar, M. Fulham, D. Feng, «Dermoscopic Image Segmentation via Multi-Stage Fully Convolutional Networks», IEEE International Symposium on Biomedical Imaging (ISBI), Venice, Italia, 2019.

[5] M. Gessert, M. Nielsen, A. Shaikh, A. Werner, R. Spreckelsen, A. Schlaefer, «Skin Lesion Diagnosis Using Ensembles, Unscaled Multi-Crop Evaluation and Loss Weighting», arXiv:2003.01048, 2020.

[6] International Skin Imaging Collaboration (ISIC), «ISIC Archive», [Online]. Disponible en: https://www.isic-archive.com/

[7] R. Azad, M. Asadi-Aghbolaghi, M. Fathy, S. Escalera, «Attention Deeplabv3+: Multi-level Context Attention Mechanism for Skin Lesion Segmentation», arXiv:2001.11242, 2020.

[8] Z. Mirikharaji, et al., «A Survey on Deep Learning for Skin Lesion Segmentation», arXiv:2206.00356, 2022.

[9] M. Rashid, et al., «Melanoma Skin Cancer Detection Using Mask-RCNN with Modified GRU», PLOS ONE, Vol. 18, No. 1, 2023.

[10] M. Goyal, et al., «Skin Lesion Segmentation and Classification Using Deep Learning and Hybrid Techniques: A Review», Diagnostics, Vol. 13, No. 19, 2023.

[11] A. V. M. Rodrigues, R. B. Oliveira, «Segmentation of Skin Lesions and Their Attributes in Dermatoscopic Images Based on Convolutional Neural Networks», Journal of Biomedical and Health Informatics, Vol. 26, No. 2, pp. 523–532, 2025.

[12] Y. Xie, J. Zhang, Y. Xia, C. Shen, «A Mutual Bootstrapping Model for Automated Skin Lesion Segmentation and Classification», arXiv:1903.03313, 2019.

[13] G. J. Chowdary, et al., «Automated Skin Lesion Segmentation Using Multi-scale Feature Extraction Scheme and Dual-attention Mechanism», arXiv:2111.08708, 2021.

[14] S. Innani, et al., «Deep Learning Based Novel Cascaded Approach for Skin Lesion Analysis», arXiv:2301.06226, 2023.

[15] O. Akinrinade, C. Du, «Skin Cancer Detection Using Deep Machine Learning Techniques», Intelligence-Based Medicine, Vol. 11, Art. 100191, 2025.

[16] Y. Wu, et al., «Detectron2», Facebook AI Research, GitHub, 2019. Disponible en: https://github.com/facebookresearch/detectron2

[17] Pavel Yakubovskiy, «Segmentation Models for PyTorch», GitHub, 2019–2022. Disponible en: https://github.com/qubvel/segmentation_models.pytorch

[18] I. Goodfellow, Y. Bengio, A. Courville, «Deep Learning», MIT Press, 2016. Disponible en: http://www.deeplearningbook.org

[19] A. V. M. Rodrigues, R. B. Oliveira, «Segmentation of Skin Lesions and Their Attributes in Dermatoscopic Images Based on Convolutional Neural Networks», *Revista de Informática Teórica e Aplicada (RITA)*, Vol. 32, No. 1, pp. 99–106, 2025. Disponible en: https://seer.ufrgs.br/index.php/rita/article/view/143546

[20] O. Ronneberger, P. Fischer, T. Brox, «U-Net: Convolutional Networks for Biomedical Image Segmentation», en: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. Wells, A.

Frangi (eds), Lecture Notes in Computer Science, vol. 9351, Springer, Cham, 2015. Disponible en: https://doi.org/10.1007/978-3-319-24574-4_28

[21] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, G. Asari, «Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation», arXiv:1802.06955, 2018.

[22] S. Bakas, et al., «Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge», arXiv:1811.02629, 2018.

[23] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, «PraNet: Parallel Reverse Attention Network for Polyp Segmentation», International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2020.

[24] G. Litjens, T. Kooi, B. E. Bejnordi, et al., «A Survey on Deep Learning in Medical Image Analysis», Medical Image Analysis, Vol. 42, pp. 60–88, 2017.

[25] D. Hendrycks, T. Dietterich, «Benchmarking Neural Network Robustness to Common Corruptions and Perturbations», International Conference on Learning Representations (ICLR), 2019.

[26] D. J. Withey, Z. J. Koles, «A Review of Medical Image Segmentation: Methods and Available Software», International Conference on Signal Processing and Communication Systems, 2008.

[27] N. Sharma, L. M. Aggarwal, «Automated Medical Image Segmentation Techniques», Journal of Medical Physics, Vol. 35, No. 1, pp. 3–14, 2010.

[28] M. Kass, A. Witkin, D. Terzopoulos, «Snakes: Active Contour Models», International Journal of Computer Vision, Vol. 1, No. 4, pp. 321–331, 1988.

[29] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, J. Liang, «UNet++: A Nested U-Net Architecture for Medical Image Segmentation», Deep Learning in Medical Image Analysis, pp. 3–11, 2018.

[30] I. Bankman, «Handbook of Medical Image Processing and Analysis», 2da Edición, Academic Press, 2008.

[31] S. Dodge, L. Karam, «A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions», 26th International Conference on Computer Communication and Networks (ICCCN), 2017.

[32] A. Skourt, A. El Hassani, A. Majda, «Lung Nodule Segmentation Using Deep Learning», Procedia Computer Science, Vol. 127, pp. 485–493, 2018. Disponible en: https://doi.org/10.1016/j.procs.2018.04.105

Appendices

Apéndice A

ANEXO: Código

Listing A.1: AAA