COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA MATEMÁTICA E INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

Análisis e interpretación de biopsias renales mediante técnicas de visión por ordenador y deep learning

Autor: Matteo Ferrari Marín

Director: David Contreras Bárcena

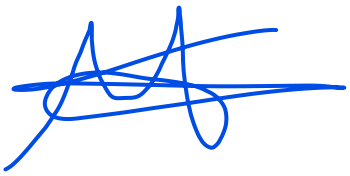
Madrid, 15 de junio de 2026

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

Análisis e interpretación de biopsias renales mediante técnicas de visión por ordenador y deep learning

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el curso académico 2025/2026 es de mi autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos.

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



Fdo.: Matteo Ferrari Marín Fecha: 15 / 06 / 2026

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: David Contreras Bárcena Fecha: 15 / 06 / 2026

EL CODIRECTOR DEL PROYECTO (si aplica)

Fdo.: No aplica Fecha: 15 / 06 / 2026

Declaración de originalidad

Declaro bajo mi responsabilidad que el Proyecto presentado con el título **Análisis e interpretación de biopsias renales mediante técnicas de visión por ordenador y deep learning** en la Escuela Técnica Superior de Ingeniería ICAI de la Universidad Pontificia Comillas en el curso académico 2025/2026 es de mi autoría y no ha sido presentado anteriormente para otros fines. El Proyecto no ha sido plagiado de ningún otro, ni total ni parcialmente, y la información que ha sido tomada de otros documentos está debidamente referenciada.

Uso de Inteligencia Artificial[1]

Declaro bajo mi responsabilidad (indicar la opción correcta):

☐ No he utilizado Inteligencia Artificial en la elaboración de este documento.

☒ He utilizado Inteligencia Artificial en la elaboración de este documento y/o del Anexo B bajo las condiciones permitidas por la Universidad Pontificia Comillas, es decir, aplicando el Nivel 2 de la Escala de Evaluación de Perkins et al. (2024): *"La IA puede utilizarse para actividades previas a la tarea, como lluvia de ideas, descripción e investigación inicial. Este nivel se centra en el uso de la IA para planificar, sintetizar y generar ideas, pero las evaluaciones deben enfatizar la capacidad de desarrollar y perfeccionar estas ideas de forma independiente"*. En concreto, la Inteligencia Artificial se ha utilizado para:

La Inteligencia Artificial se ha empleado con tres finalidades concretas: (1) dar formato en LaTeX a los contenidos de la memoria, partiendo de borradores en texto plano; (2) depurar el código de los notebooks de entrenamiento e inferencia, identificando errores y sugiriendo correcciones puntuales; y (3) apoyar la búsqueda y selección de artículos científicos relacionados con el tema del proyecto, facilitando la identificación de trabajos relevantes en el ámbito del *Multiple Instance Learning* y la histopatología computacional.

[1] Esta declaración se refiere al uso de Inteligencia Artificial generativa para la elaboración de los documentos del Proyecto (Anexo B y Memoria). No se aplica a Proyectos en los que, por su naturaleza, deba utilizarse inteligencia artificial como parte de los mismos (aplicación de técnicas de aprendizaje automático, redes neuronales, análisis de datos...).

 (firmar aquí)
Firma: Matteo Ferrari Marín
Fecha: 15 / 06 / 2026

Autorización para la entrega del Proyecto

Director del TFG
 (firmar aquí)
Firma: David Contreras Bárcena
Fecha: 15 / 06 / 2026

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA MATEMÁTICA E INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

Análisis e interpretación de biopsias renales mediante técnicas de visión por ordenador y deep learning

Autor: Matteo Ferrari Marín

Director: David Contreras Bárcena

Madrid, 15 de junio de 2026

Agradecimientos

En primer lugar, quiero expresar mi gratitud al Doctor Rafael Marín Iranzo, por su supervisión, orientación y dedicación a lo largo de todos estos meses de trabajo. Su apoyo ha sido fundamental para llevar este proyecto adelante.

Quiero agradecer también a mis compañeros, que me han acompañado durante todo el grado y han hecho de este camino una experiencia mucho más enriquecedora. Y, por encima de todo, a mi familia, por su apoyo incondicional en cada etapa.

ANÁLISIS E INTERPRETACIÓN DE BIOPSIAS RENALES MEDIANTE TÉCNICAS DE VISIÓN POR ORDENADOR Y DEEP LEARNING

Autor: Matteo Ferrari Marín
Director: David Contreras Bárcena

Resumen

Se propone un sistema de clasificación automática de biopsias renales basado en *Multiple Instance Learning* (MIL) sobre *Whole Slide Images* procedentes del *Kidney Precision Medicine Project* (KPMP) [1]. **Este trabajo constituye, hasta donde se conoce, el primer estudio que aborda la clasificación simultánea de cuatro categorías diagnósticas sobre WSIs del KPMP**: enfermedad renal crónica (CKD), lesión renal aguda (AKI), nefropatía diabética (DM-R) y tejido sano de referencia (*Healthy Reference*). Combinando el modelo fundacional H-optimus-0 [2], la arquitectura TransformerMIL [3] y *Test-Time Augmentation*, el sistema alcanza un F1-macro de 0,933 y una exactitud del 92,5 % sobre 106 biopsias de test, con explicabilidad clínicamente validada en tres niveles.

Palabras clave: biopsias renales; *Multiple Instance Learning*; *Whole Slide Images*; KPMP; modelos fundacionales; TransformerMIL; H-optimus; explicabilidad; atención.

Resumen ejecutivo

1 Introducción

Las enfermedades renales crónicas y agudas representan uno de los principales problemas de salud pública a nivel mundial, afectando a más de 850 millones de personas. Su diagnóstico definitivo requiere la práctica de una biopsia renal y la posterior valoración histológica por parte de un patólogo especializado. Este proceso es lento (horas–días por caso), costoso y sujeto a una variabilidad interobservador significativa, especialmente para las etiologías más complejas como la enfermedad renal crónica (CKD) o la nefropatía diabética (DM-R).

La histopatología computacional ofrece una oportunidad única para automatizar y estandarizar este análisis. Sin embargo, las imágenes de portaobjetos completos (*Whole Slide Images*, WSI) presentan un reto particular: son enormes (hasta 100.000×100.000 píxeles) y carecen de anotaciones a nivel de píxel, ya que los diagnósticos están disponibles únicamente

al nivel del portaobjetos. El paradigma *Multiple Instance Learning* (MIL) [4] resuelve este problema de forma elegante: la WSI se divide en pequeños parches (*patches*) que constituyen instancias de una bolsa (*bag*), y el modelo aprende a clasificar la bolsa entera a partir de los parches individuales, sin necesidad de anotaciones por parche.

2 Objetivos

El objetivo principal de este trabajo es diseñar, implementar y evaluar un sistema automático de clasificación de biopsias renales mediante MIL, capaz de discriminar cuatro categorías diagnósticas: CKD, AKI, DM-R y tejido sano de referencia.

Los objetivos específicos son: (1) construir un pipeline de preprocesamiento robusto para WSIs, incluyendo tiling adaptativo y filtrado de fondo; (2) comparar cinco *backbones* para la extracción de características, desde ResNet50 hasta modelos fundacionales de histopatología; (3) evaluar dos arquitecturas de agregación MIL (ABMIL y TransformerMIL); (4) implementar tres niveles de explicabilidad (XAI) coherentes con el conocimiento clínico.

3 Descripción del sistema

El sistema implementado sigue un pipeline en cuatro etapas:

1. Preprocesamiento. Cada WSI se divide en parches de 256×256 píxeles a 20× de aumento. Los parches con más de un 40 % de fondo se descartan mediante un filtro colorimétrico en espacio HSV, complementado con un filtro de nitidez basado en la varianza del Laplaciano (umbral: 45), lo que permite recuperar parches límitrofes rechazados erróneamente por el filtro de color.

2. Extracción de características. Se comparan cinco extractores: ResNet50 (preentrenado en ImageNet), CONCH, UNI, Phikon y H-optimus-0 [2]. El extractor final, H-optimus-0, es un Vision Transformer (ViT-g/14) con 1.100M de parámetros preentrenado mediante autosupervisión DINOv2 sobre más de 500.000 imágenes histológicas de alta resolución. Produce embeddings de 1.536 dimensiones por parche.

3. Agregación MIL. Se implementan ABMIL (*Attention-Based MIL*) y TransformerMIL [3], que aplica un Transformer de dos capas sobre el conjunto de parches, capturando interacciones espaciales entre regiones. El sistema final usa TransformerMIL, que supera a ABMIL en 4,9 puntos de F1-macro con H-optimus-0.

4. Test-Time Augmentation (TTA). Durante la inferencia se realizan $K = 5$ submuestreos aleatorios del conjunto de parches y se promedian las distribuciones de probabilidad, estabilizando la predicción y mejorando el F1-macro en 0,008 puntos.

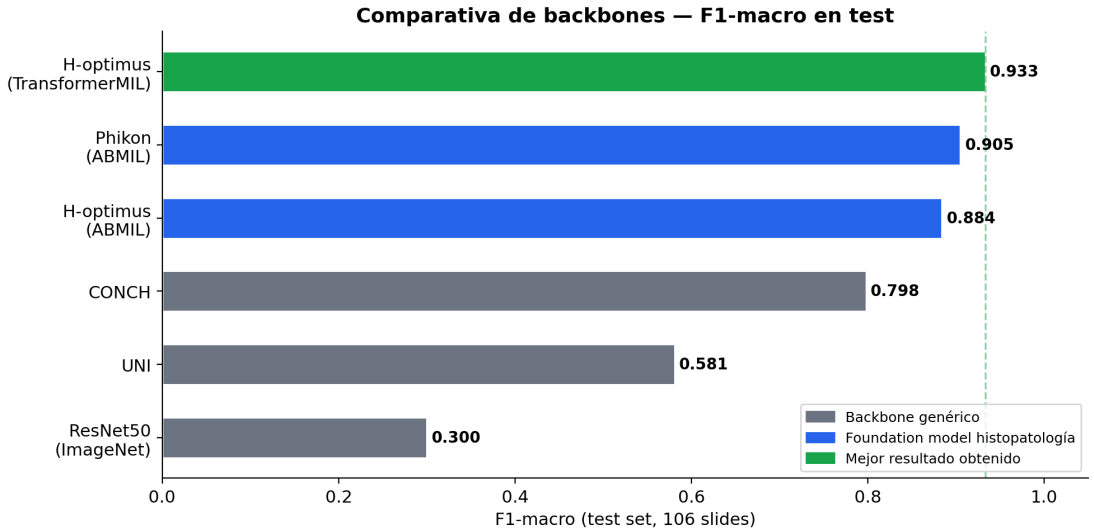


Figura 1: Comparativa de F1-macro en test (106 slides) para los distintos backbones evaluados con arquitectura ABMIL.

La Figura 1 ilustra el impacto del backbone. La diferencia entre ResNet50 (F1≈0,30) y H-optimus-0 con TransformerMIL (F1=0,933) es de 47 puntos porcentuales, evidenciando que el preentrenamiento autosupervisado sobre imágenes histológicas es el factor más determinante del rendimiento.

4 Resultados

El sistema final (H-optimus-0 + TransformerMIL + TTA $K = 5$) se evalúa sobre 106 biopsias de test:

Clase	**Precision**	**Recall**	**F1**	**N**
CKD	0,87	0,90	0,88	29
AKI	0,96	0,90	0,93	30
DM-R	0,90	0,93	0,92	30
Healthy	1,00	1,00	1,00	17
Macro avg	**0,93**	**0,93**	**0,933**	106

La clase *Healthy Reference* se clasifica con precisión perfecta (F1=1,00). CKD es la clase más difícil (F1=0,88), consecuencia de su heterogeneidad histológica y del solapamiento con estadios avanzados de DM-R. Los tres niveles de explicabilidad producen visualizaciones clínicamente plausibles, validadas por un especialista en nefrología: focos de fibrosis intersticial en CKD, daño tubular agudo en AKI y nódulos de Kimmelstiel-Wilson en DM-R.

5 Conclusiones

Los resultados permiten extraer cuatro conclusiones: (i) los modelos fundacionales de histopatología son indispensables, puesto que ResNet50 fracasa en este dominio (F1≈0,30); (ii) la arquitectura de agregación importa cuando el backbone proporciona embeddings de alta dimensión (TransformerMIL supera a ABMIL en 4,9 puntos con H-optimus-0); (iii) el TTA con $K = 5$ estabiliza y mejora la inferencia sin coste de reentrenamiento; y (iv) la explicabilidad multi-nivel es coherente con el conocimiento clínico nefropatológico.

UNIVERSIDAD PONTIFICIA COMILLAS
Escuela Técnica Superior de Ingeniería (ICAI)
Grado en Ingeniería Matemática e Inteligencia Artificial

ANALYSIS AND INTERPRETATION OF RENAL BIOPSIES USING COMPUTER VISION AND DEEP LEARNING TECHNIQUES

Author: Matteo Ferrari Marín
Director: David Contreras Bárcena

Abstract

An automated classification system for renal biopsies based on Multiple Instance Learning (MIL) applied to Whole Slide Images from the Kidney Precision Medicine Project (KPMP) [1] is proposed. **To the best of our knowledge, this is the first study to simultaneously classify four diagnostic categories on KPMP WSIs**: Chronic Kidney Disease (CKD), Acute Kidney Injury (AKI), Diabetic Mellitus-Related nephropathy (DM-R), and Healthy Reference tissue. By combining the H-optimus-0 foundation model [2], TransformerMIL aggregation [3], and Test-Time Augmentation, the system achieves a macro F1-score of 0.933 and 92.5% accuracy on 106 test slides, with clinically validated explainability at three levels.

Keywords: renal biopsy; Multiple Instance Learning; Whole Slide Images; KPMP; foundation models; TransformerMIL; H-optimus; explainability; attention.

Executive Summary

1 Introduction

Chronic and acute kidney diseases are among the leading public health challenges worldwide, affecting more than 850 million people. Definitive diagnosis requires renal biopsy followed by histological assessment by a specialist pathologist, a slow and costly process subject to significant inter-observer variability, especially for complex etiologies such as CKD or diabetic nephropathy (DM-R).

Computational histopathology offers a unique opportunity to automate and standardise this analysis. However, Whole Slide Images (WSIs) pose a specific challenge: they are enormous (up to 100,000×100,000 pixels) and lack pixel-level annotations, as diagnoses are only available at slide level. The Multiple Instance Learning (MIL) paradigm [4] elegantly addresses this: the WSI is split into small patches constituting instances of a bag, and the model learns to classify the entire bag from individual patches, without patch-level annotations.

2 Objectives

The main objective is to design, implement, and evaluate an automated renal biopsy classification system using MIL, capable of discriminating four diagnostic categories: CKD, AKI, DM-R, and Healthy Reference tissue.

Specific objectives are: (1) build a robust WSI preprocessing pipeline; (2) compare five feature extraction backbones, from ResNet50 to histopathology foundation models; (3) evaluate two MIL aggregation architectures (ABMIL and TransformerMIL); (4) implement three levels of explainability (XAI) consistent with clinical knowledge.

3 System Description

The implemented system follows a four-stage pipeline:

1. Preprocessing. Each WSI is divided into 256×256 pixel patches at 20× magnification. Patches with more than 40% background are discarded using an HSV colorimetric filter complemented by a Laplacian variance sharpness filter (threshold: 45).

2. Feature extraction. Five extractors are compared: ResNet50, CONCH, UNI, Phikon, and H-optimus-0 [2]. The final extractor, H-optimus-0, is a ViT-G/14 with 1.1B parameters pre-trained via DINOv2 on over 500,000 histological images, producing 1,536-dimensional embeddings per patch.

3. MIL Aggregation. ABMIL and TransformerMIL [3] are implemented. TransformerMIL applies a two-layer Transformer over the bag's patch set, capturing spatial interactions. It outperforms ABMIL by 4.9 F1 points with H-optimus-0.

4. Test-Time Augmentation (TTA). $K = 5$ random subsamplings are averaged during inference, stabilising predictions and improving macro F1 by 0.008 points.

4 Results

The final system achieves F1=0.933 and Acc=92.5% on 106 test biopsies. Healthy Reference is classified with perfect accuracy (F1=1.00). The three XAI levels produce clinically plausible visualisations validated by a nephrology specialist: interstitial fibrosis in CKD, acute tubular damage in AKI, and Kimmelstiel-Wilson nodules in DM-R.

5 Conclusions

(i) Histopathology foundation models are indispensable (ResNet50 fails, F1≈0.30); (ii) the aggregation architecture matters with high-dimensional embeddings (TransformerMIL outperforms ABMIL by 4.9 points); (iii) TTA stabilises inference at no retraining cost; (iv) the multi-level explainability is consistent with clinical nephropathological knowledge.

Índice

Lista de acrónimos **15**

Capítulo 1 Introducción **16**

1.1 Motivación y contexto clínico 16
1.2 Objetivos del trabajo 17
1.3 Alcance y limitaciones 18
1.4 Alineación con los Objetivos de Desarrollo Sostenible 18
1.5 Estructura del documento 19

Capítulo 2 Estado del arte **20**

2.1 Patología digital y Whole Slide Images 20
2.2 Deep learning en histopatología 20
2.3 Modelos fundacionales para histopatología 21
2.3.1 Vision Transformer como extractor de características 21
2.3.2 CONCH 21
2.3.3 Phikon 22
2.3.4 H-optimus-0 22
2.4 Multiple Instance Learning 23
2.4.1 Formulación del problema 23
2.4.2 AttentionMIL (ABMIL) 23
2.4.3 CLAM-SB 24
2.4.4 TransMIL 24
2.4.5 TransformerMIL 25
2.5 Explicabilidad en modelos de visión (XAI) 25
2.5.1 Attention maps 26
2.5.2 Grad-CAM 26
2.5.3 ViT attention maps 27
2.6 Trabajos previos sobre el dataset KPMP 27

Capítulo 3 Dataset **29**

3.1 Kidney Precision Medicine Project (KPMP) 29
3.2 Clases de diagnóstico 29
3.3 Tinciones histológicas: PAS y Trichrome 30

3.4 Distribución del dataset y desbalance de clases . . . 31
3.5 División train / validación / test . . . 32

Capítulo 4 Metodología 33
4.1 Visión general del pipeline . . . 33
4.2 Preprocesamiento de WSI . . . 33
4.2.1 Tiling con OpenSlide . . . 33
4.2.2 Filtrado de fondo: HSV y varianza del Laplaciano . . . 34
4.2.3 Caché de coordenadas . . . 35
4.3 Modelos fundacionales como extractores de características . . . 35
4.4 Arquitecturas MIL implementadas . . . 35
4.4.1 ABMIL . . . 35
4.4.2 TransformerMIL . . . 36
4.4.3 CLAM-SB . . . 37
4.4.4 TransMIL . . . 37
4.5 Estrategia de entrenamiento . . . 38
4.5.1 Función de pérdida ponderada . . . 38
4.5.2 Optimizador y scheduler . . . 39
4.5.3 Early stopping . . . 39
4.5.4 Submuestreo de parches . . . 39
4.6 Test-Time Augmentation (TTA) . . . 39
4.7 Módulos de explicabilidad (XAI) . . . 40
4.7.1 Nivel 1 — Mapa de calor de atención sobre la WSI . . . 40
4.7.2 Nivel 2A — Visualización de los parches más relevantes . . . 40
4.7.3 Nivel 2B — Atención interna del ViT (register-aware) . . . 41

Capítulo 5 Experimentos y resultados 42
5.1 Configuración experimental . . . 42
5.2 Experimento 1: Comparativa de backbones con ABMIL . . . 42
5.3 Experimento 2: Comparativa de arquitecturas MIL con Phikon y H-optimus 43
5.3.1 Arquitecturas MIL de referencia: CLAM-SB y TransMIL . . . 45
5.3.2 Curvas de entrenamiento: H-optimus + TransformerMIL . . . 45
5.4 Experimento 3: Test-Time Augmentation . . . 47
5.5 Resultados del sistema final . . . 48

5.6 Análisis de explicabilidad 51
5.6.1 CKD — Enfermedad renal crónica 52
5.6.2 AKI — Lesión renal aguda 53
5.6.3 DM-R — Nefropatía diabética 54
5.6.4 Healthy — Tejido renal sano 55
5.6.5 Discusión clínica de los resultados XAI 55
5.7 Discusión de resultados 56
5.7.1 Impacto del backbone 56
5.7.2 *Overfitting* y regularización 56
5.7.3 Clase Healthy: clasificación perfecta 57
5.7.4 Clase CKD: mayor dificultad de clasificación 57

Capítulo 6 Conclusiones y trabajo futuro 58
6.1 Conclusiones 58
6.2 Limitaciones 59
6.3 Líneas de trabajo futuro 60

Bibliografía 62

Índice de figuras

1 Distribución de biopsias por clase diagnóstica en el dataset KPMP utilizado. Se indica entre paréntesis el número de slides en el conjunto de test para cada clase. 31
2 Comparativa de F1-macro en test para los distintos *backbones* evaluados con arquitectura ABMIL. Los modelos fundacionales de histopatología superan ampliamente a ResNet50 preentrenado en ImageNet. 43
3 F1-macro y Accuracy en test para las combinaciones de *backbone* y arquitectura MIL evaluadas. TransformerMIL supera a ABMIL únicamente con H-optimus-0, cuya mayor dimensionalidad de *embedding* aprovecha el mecanismo de self-attention entre parches. 44
4 Curvas de entrenamiento del modelo H-optimus + TransformerMIL. La pérdida de entrenamiento decrece de 1,14 a 0,06, mientras el F1 de validación alcanza su máximo en la época 35 (0,921) y se mantiene ruidoso por el pequeño tamaño del conjunto de validación. 46
5 Efecto del número de pasadas TTA (K) sobre el F1-macro en test. Con $K = 5$ ya se alcanza el rendimiento máximo y se elimina la varianza debida al submuestreo aleatorio de parches. 47
6 Precision, Recall y F1-score por clase diagnóstica del sistema final (H-optimus + TransformerMIL + TTA $K = 5$). La clase *Healthy* se clasifica con perfección absoluta; CKD presenta la mayor dificultad. 49
7 Matriz de confusión del modelo H-optimus-0 + ABMIL (F1-macro=0,884, Accuracy=0,877). Los principales errores se producen en CKD (6 errores: 4 confundidos con DM-R) y en DM-R (4 errores: 3 confundidos con CKD). . . 50
8 Matriz de confusión del sistema final H-optimus-0 + TransformerMIL + TTA (F1-macro=0,933, Accuracy=0,925). TransformerMIL reduce los errores de 13 a 8, principalmente mejorando la frontera entre CKD y DM-R. *Healthy* se clasifica sin errores. 51

9 Parches de mayor atención MIL (fila superior) y mapas de atención interna del ViT H-optimus-0 (fila inferior) para una biopsia CKD clasificada con confianza=1,000. Tinción PAS. Los tres parches presentan atención MIL máxima (1,000). La atención interna del ViT (amarillo-naranja) se concentra en las zonas de expansión intersticial y en la interfaz entre el tejido fibrótico y los túbulos atróficos. 52

10 Parches de mayor atención MIL y mapas de atención interna del ViT para una biopsia AKI con confianza=1,000. Tinción PAS. La atención MIL desciende ligeramente entre parches (1,000, 0,981, 0,975), reflejando mayor selectividad que en CKD. La atención interna del ViT se concentra en áreas con detritos celulares intraluminales y epitelio tubular dañado. 53

11 Parches de mayor atención MIL y mapas de atención interna del ViT para una biopsia DM-R con confianza=1,000. Tinción Trichrome (tricrómico de Masson): azul = colágeno/fibrosis, rojo = citoplasma, negro = núcleos. La atención interna del ViT se distribuye sobre las estructuras glomerulares con expansión mesangial y depósitos nodulares de colágeno. 54

12 Parches de mayor atención MIL y mapas de atención interna del ViT para una biopsia Healthy con confianza=1,000. Tinción PAS. La atención MIL es alta (1,000, 0,995, 0,994) pero la atención interna del ViT presenta una distribución más difusa que en las clases patológicas, coherente con la ausencia de hallazgos focales. 55

Índice de tablas

1 Distribución de biopsias por clase diagnóstica en el dataset KPMP utilizado. 31
2 Distribución de biopsias por partición del dataset. 32
3 Distribución del conjunto de test por clase. 32
4 Hiperparámetros comunes a todos los experimentos. 42
5 Comparativa de *backbones* con arquitectura ABMIL en el test set (106 slides). 42
6 Comparativa de arquitecturas MIL con Phikon y H-optimus en el test set. . . 44
7 Comparativa de F1-macro en test (106 slides) entre las cuatro arquitecturas evaluadas. 45
8 Efecto del TTA sobre el F1-macro en test (K pasadas de submuestreo). . . . 47
9 Informe de clasificación del sistema final (H-optimus + TransformerMIL + TTA $K = 5$). 48

Lista de acrónimos

AKI	Acute Kidney Injury (lesión renal aguda)
ABMIL	Attention-Based Multiple Instance Learning
AUC	Area Under the ROC Curve
CKD	Chronic Kidney Disease (enfermedad renal crónica)
CNN	Convolutional Neural Network
DM-R	Diabetes Mellitus-Related kidney disease (nefropatía diabética)
GPU	Graphics Processing Unit
HSV	Hue-Saturation-Value (espacio de color)
IFTA	Interstitial Fibrosis and Tubular Atrophy
KPMP	Kidney Precision Medicine Project
MIL	Multiple Instance Learning
MLP	Multi-Layer Perceptron
PAS	Periodic Acid–Schiff (tinción histológica)
SSL	Self-Supervised Learning
TTA	Test-Time Augmentation
ViT	Vision Transformer
WSI	Whole Slide Image (imagen de portaobjetos completo)
XAI	Explainable Artificial Intelligence

Capítulo 1 Introducción

1.1. Motivación y contexto clínico

Las enfermedades renales constituyen uno de los problemas de salud pública más prevalentes del siglo XXI. Según estimaciones recientes, más de 850 millones de personas en el mundo padecen algún tipo de enfermedad renal, y se prevé que la enfermedad renal crónica (*Chronic Kidney Disease*, CKD) se convierta en la quinta causa de muerte prematura a nivel global para el año 2040 [5]. A esta carga se añaden la lesión renal aguda (*Acute Kidney Injury*, AKI), asociada a elevadas tasas de mortalidad hospitalaria, y la nefropatía diabética (*Diabetic Mellitus-Related*, DM-R), complicación frecuente de la diabetes mellitus tipo 2 y primera causa de insuficiencia renal terminal en los países desarrollados [6].

El diagnóstico definitivo de estas patologías requiere la obtención de una biopsia renal, cuyo análisis histológico permite al patólogo identificar lesiones estructurales como la fibrosis intersticial y atrofia tubular (IFTA), los cambios glomerulares o la inflamación. Este proceso se realiza sobre imágenes de microscopía óptica obtenidas tras la tinción de los tejidos con reactivos específicos —entre los más habituales, el ácido peryódico de Schiff (PAS) y el tricrómico de Masson (Trichrome)— que realzan diferentes componentes celulares y extracelulares.

Sin embargo, la interpretación de las biopsias renales presenta importantes desafíos. En primer lugar, la valoración histológica es un proceso manual intensivo que exige la formación especializada de anatomopatólogos con experiencia en patología renal, profesionales escasos en muchos sistemas sanitarios. En segundo lugar, la decisión diagnóstica está sujeta a variabilidad interobservador: diferentes expertos pueden asignar interpretaciones distintas a una misma biopsia, especialmente en casos límite. Por último, los avances en los sistemas de digitalización han dado lugar a las imágenes de portaobjetos completos (*Whole Slide Images*, WSI), archivos de alta resolución que pueden superar los 100.000 $\times$ 100.000 píxeles, y cuyo análisis manual resulta extraordinariamente lento.

En este contexto, la inteligencia artificial y, en particular, el aprendizaje profundo (*deep learning*), ofrecen una oportunidad para automatizar y estandarizar el análisis de biopsias renales. El desarrollo de sistemas de ayuda al diagnóstico basados en visión por ordenador podría reducir la carga de trabajo del patólogo, aumentar la reproducibilidad de los diagnósticos y, en última instancia, mejorar los resultados clínicos de los pacientes.

1.2. Objetivos del trabajo

El presente Trabajo de Fin de Grado tiene como objetivo principal el diseño, implementación y evaluación de un sistema de clasificación automática de biopsias renales mediante técnicas de aprendizaje profundo aplicadas a imágenes WSI del *Kidney Precision Medicine Project* (KPMP).

De forma más concreta, los objetivos específicos son los siguientes:

1. **Construcción del pipeline de preprocesamiento.** Desarrollar un sistema de extracción de parches (*tiling*) sobre WSIs que permita su procesamiento eficiente sin necesidad de cargar la imagen completa en memoria, incorporando filtrado automático de fondo mediante criterios colorimétricos (HSV) y de nitidez (varianza del Laplaciano).

2. **Evaluación de modelos fundacionales para histopatología.** Comparar cinco extractores de características —ResNet50 [7], CONCH [8], UNI [9], Phikon [10] y H-optimus [2]— en términos de la calidad de los embeddings para la tarea de clasificación de biopsias renales.

3. **Implementación y comparación de arquitecturas MIL.** Implementar dos arquitecturas de *Multiple Instance Learning*: ABMIL [11] y TransformerMIL, y comparar su rendimiento en la clasificación de las WSIs en cuatro categorías diagnósticas: CKD, AKI, DM-R y tejido renal sano (*Healthy Reference*).

4. **Mejora mediante Test-Time Augmentation.** Investigar el uso de TTA (*Test-Time Augmentation*) sobre submuestras de parches para estabilizar las predicciones y mejorar la robustez del sistema final.

5. **Explicabilidad del modelo (XAI).** Implementar tres niveles de explicabilidad: (i) mapas de calor de atención que identifican los parches más relevantes dentro de la WSI, (ii) visualización de los parches de mayor atención en RGB, y (iii) análisis de la atención interna del ViT del backbone (*register-aware ViT attention*) para localizar estructuras histológicas dentro de cada parche.

6. **Análisis de los resultados desde una perspectiva clínica.** Relacionar las regiones de alta atención identificadas por el modelo con las lesiones histológicas conocidas de cada patología, contrastando los hallazgos con criterio nefrológico.

1.3. Alcance y limitaciones

El sistema desarrollado en este trabajo se enmarca dentro de las siguientes consideraciones:

Alcance:

- Se trabaja exclusivamente con datos de acceso abierto del portal KPMP (*Open Access*), sin necesidad de acuerdo de uso de datos (*Data Use Agreement*, DUA).
- Las tinciones consideradas son PAS y Trichrome, las más informativas para la valoración de fibrosis renal y daño glomerular.
- El entrenamiento se realiza en Google Colab Pro, haciendo el trabajo reproducible sin infraestructura computacional dedicada.
- La arquitectura TransformerMIL propuesta es una solución original que combina un mecanismo de atención basado en *token* CLS con un encoder Transformer, adaptado específicamente al dominio de las WSIs con modelos fundacionales.

Limitaciones:

- El número de slides disponibles en la clase *Healthy Reference* es inferior al de las otras tres clases, lo que introduce un desbalance de clases que se mitiga mediante pérdidas ponderadas.
- Los modelos se entrenan con etiquetas a nivel de slide (sin anotaciones de estructuras individuales), lo que limita la precisión de la explicabilidad espacial a nivel de parche.
- El dataset es de tamaño moderado (710 slides), lo que favorece el sobreajuste con modelos muy profundos; se mitiga mediante regularización y early stopping.
- Los resultados son preliminares y no han sido validados en un entorno clínico formal.

1.4. Alineación con los Objetivos de Desarrollo Sostenible

Este trabajo contribuye a los siguientes Objetivos de Desarrollo Sostenible de las Naciones Unidas:

- **ODS 3 — Salud y bienestar.** El sistema desarrollado contribuye directamente a mejorar la detección temprana de enfermedades renales crónicas y agudas, reduciendo el

tiempo de diagnóstico y potencialmente el coste sanitario asociado. La automatización del análisis histopatológico puede hacer el diagnóstico accesible en centros sin patólogos especializados en nefropatología.

- **ODS 10 — Reducción de las desigualdades.** La escasez de patólogos renales especializados es especialmente aguda en países de ingresos medios y bajos. Un sistema de ayuda al diagnóstico basado en inteligencia artificial podría democratizar el acceso a un diagnóstico histopatológico de calidad, reduciendo la brecha entre sistemas sanitarios con diferentes recursos.
- **ODS 9 — Industria, innovación e infraestructura.** El trabajo contribuye al desarrollo de herramientas de salud digital basadas en inteligencia artificial, enmarcadas en la transformación de la práctica médica mediante tecnologías de visión por ordenador y aprendizaje profundo.

1.5. Estructura del documento

El resto del documento se organiza de la siguiente manera:

- El **Capítulo 2** revisa el estado del arte en patología digital computacional, *deep learning* aplicado a histopatología, los modelos fundacionales específicos de histología, las principales arquitecturas MIL y los métodos de explicabilidad relevantes.
- El **Capítulo 3** describe el dataset KPMP utilizado, incluyendo las características de las imágenes, la distribución de las cuatro clases diagnósticas, las tinciones histológicas y la división del conjunto de datos.
- El **Capítulo 4** detalla el diseño completo del sistema: pipeline de preprocesamiento, modelos fundacionales como extractores de características, arquitecturas MIL implementadas, estrategia de entrenamiento, TTA y módulos de explicabilidad.
- El **Capítulo 5** presenta los experimentos realizados, los resultados obtenidos, la comparativa entre *backbones* y arquitecturas MIL, y el análisis de explicabilidad con interpretación clínica.
- El **Capítulo 6** recoge las conclusiones del trabajo, sus limitaciones y las líneas de trabajo futuro.

Capítulo 2 Estado del arte

2.1. Patología digital y Whole Slide Images

La patología digital surge de la convergencia entre los escáneres de diapositivas de alta resolución y el almacenamiento digital masivo, permitiendo digitalizar por completo los portaobjetos histológicos tradicionales. El resultado son las *Whole Slide Images* (WSI), imágenes de dimensiones que típicamente oscilan entre 50.000 × 50.000 y 150.000 × 150.000 píxeles, con una resolución espacial de aproximadamente 0,25 μm por píxel al nivel de mayor aumento (40×).

Las WSI se almacenan en formatos piramidales multiresolución —los más comunes son `.svs` (Aperio), `.ndpi` (Hamamatsu) y `.tif` (TIFF genérico)— que permiten acceder eficientemente a regiones de interés a distintos niveles de zoom sin cargar la imagen completa en memoria. La librería OpenSlide [12] es el estándar de facto para la lectura de estos formatos en entornos de investigación.

La digitalización ha impulsado el campo de la patología computacional (*computational pathology*), que aplica técnicas de visión por ordenador e inteligencia artificial para asistir al patólogo en tareas como la detección de tumores, la cuantificación de marcadores y la estratificación del riesgo [13]. Sin embargo, el enorme tamaño de las WSI plantea desafíos computacionales y algorítmicos específicos que han motivado el desarrollo de metodologías adaptadas.

2.2. Deep learning en histopatología

Las redes neuronales convolucionales (CNN) han demostrado capacidades sobresalientes en el reconocimiento de patrones en imágenes histológicas. Los primeros trabajos relevantes en patología digital aplicaban CNNs a parches extraídos de las WSI para tareas de clasificación a nivel de parche [13], requiriendo anotaciones a nivel de píxel o región, que son costosas de obtener y escasas en la práctica clínica.

Un hito fundamental fue el trabajo de Campanella et al. [14], que demostró que modelos de aprendizaje débilmente supervisado entrenados con más de 44.000 WSIs podían alcanzar un rendimiento diagnóstico equiparable al del patólogo experto en varios tipos de cáncer, utilizando únicamente etiquetas a nivel de portaobjetos. Este resultado motivó un cambio de paradigma hacia métodos que aprovechan etiquetas débiles, más fáciles de obtener.

Más recientemente, la irrupción de los *vision transformers* y los modelos fundacionales (*foundation models*) entrenados sobre millones de parches histológicos ha abierto nuevas perspectivas, al ofrecer representaciones universales de alta calidad que mejoran el rendimiento en tareas *downstream* con pocos datos etiquetados. Estos modelos se describen en detalle en la Sección 2.3.

2.3. Modelos fundacionales para histopatología

Los modelos fundacionales (*foundation models*) son modelos de aprendizaje profundo entrenados mediante aprendizaje autosupervisado (*self-supervised learning*, SSL) sobre enormes conjuntos de datos sin etiquetas. En este contexto, el término *backbone* designa la red neuronal base encargada de transformar cada parche de imagen en un vector de características (*embedding*); es, en esencia, el extractor de información visual sobre el que se construye el resto del sistema. En el dominio de la histopatología, los *backbones* fundacionales han demostrado proporcionar *embeddings* de alta calidad que superan significativamente a los extractores clásicos como ResNet50 [7] preentrenado en ImageNet, al haber sido entrenados específicamente sobre millones de parches de tejido histológico real.

2.3.1. Vision Transformer como extractor de características

El *Vision Transformer* (ViT), propuesto por Dosovitskiy et al. [15], adapta la arquitectura Transformer de procesamiento de lenguaje natural a la visión por ordenador dividiendo la imagen en parches regulares de tamaño fijo (por ejemplo, 16×16 píxeles) y procesándolos como secuencias de *tokens*. Un *token* especial CLS (*class token*) agrega la información global de la imagen a través de los mecanismos de *self-attention* apilados, y su representación en la capa final se usa como descriptor de la imagen.

La clave del éxito de los ViTs en histopatología radica en que el mecanismo de self-attention permite al modelo capturar dependencias de largo alcance entre estructuras histológicas distantes, algo que las CNN capturan de forma más limitada dada su naturaleza local.

2.3.2. CONCH

CONCH (*CONtrastive learning from Captions for Histopathology*) [8] es un modelo visión-lenguaje entrenado mediante aprendizaje contrastivo sobre más de 1,17 millones de pares

imagen-texto procedentes de patología. Su *backbone* visual es un ViT-B/16 entrenado conjuntamente con un codificador de texto, lo que le permite tanto clasificar parches como responder a consultas en lenguaje natural. El vector de *embedding* tiene dimensión 512.

2.3.3. Phikon

Phikon [10] es un modelo fundacional desarrollado por Owkin, entrenado mediante *Masked Image Modeling* (MIM) con el objetivo de escalado autosupervisado sobre más de 40 millones de parches de histopatología procedentes de 6.093 portaobjetos. Su arquitectura es un ViT-B/16 con 12 capas, 12 cabezas de atención y dimensión de *embedding* 768. El preentrenamiento con MIM obliga al modelo a reconstruir regiones enmascaradas de parches histológicos, forzando la extracción de características locales y globales de alta calidad.

2.3.4. H-optimus-0

H-optimus-0 [2] es el modelo fundacional de mayor escala disponible públicamente para histopatología computacional, desarrollado por Bioptimus. Está basado en ViT-G/14 (*Giant*), la variante más grande de la familia ViT, con 1.100 millones de parámetros, 40 bloques Transformer, 24 cabezas de atención por bloque y un vector de *embedding* de dimensión 1.536.

El preentrenamiento de H-optimus-0 sigue la metodología DINOv2 [16], un marco de aprendizaje autosupervisado basado en destilación de conocimiento entre redes "alumno" y "profesor", sobre una colección masiva de imágenes histológicas de diversas fuentes, modalidades de tinción y órganos. Una característica destacada de su arquitectura es la incorporación de 4 *register tokens* [16]: *tokens* adicionales entre el CLS y los *tokens* de parche cuya función es absorber los artefactos de alta norma que aparecen en los mapas de atención de los ViTs de gran escala (especialmente en regiones de fondo uniforme), mejorando así la calidad de las representaciones y la interpretabilidad de la atención.

Su dimensionalidad de *embedding* superior (1.536 frente a las 768 de Phikon) y la escala del preentrenamiento hacen de H-optimus-0 el extractor de características con mayor capacidad representacional entre los evaluados en este trabajo.

2.4. Multiple Instance Learning

2.4.1. Formulación del problema

El aprendizaje con múltiples instancias (*Multiple Instance Learning*, MIL) es un paradigma de aprendizaje supervisado débil especialmente adecuado para el análisis de WSI. Fue introducido formalmente por Dietterich et al. en el contexto de la predicción de actividad molecular, y posteriormente extendido a multitud de dominios.

En MIL, los datos se organizan en *bolsas* (*bags*), cada una compuesta por un conjunto de instancias. En el contexto de las WSI:

- **Bolsa**: una WSI completa, representada como un conjunto de N parches $\mathcal{B} = \{x_1, x_2, \ldots, x_N\}$.
- **Instancia**: un parche individual $x_i \in \mathbb{R}^{C \times H \times W}$.
- **Etiqueta**: se conoce únicamente a nivel de bolsa $Y \in \{0, 1, \ldots, K-1\}$ (diagnóstico del paciente), no a nivel de instancia.

El objetivo es aprender un clasificador $f : \mathcal{B} \to Y$ que, dada una bolsa de instancias, prediga la etiqueta de la bolsa. La hipótesis estándar de MIL establece que una bolsa es positiva si y solo si contiene al menos una instancia positiva; sin embargo, en la práctica clínica esta suposición se relaja, considerando que la etiqueta de la bolsa emerge de las relaciones entre instancias.

El reto fundamental de MIL para WSI es el agregador: cómo combinar las representaciones de los N parches (potencialmente varios miles) en una única representación de la bolsa sobre la que aplicar el clasificador.

2.4.2. AttentionMIL (ABMIL)

Ilse et al. [11] propusieron un mecanismo de agregación basado en atención que aprende a ponderar la contribución de cada instancia a la representación de la bolsa. Dado un extractor de características ϕ que mapea cada parche a un vector $\boldsymbol{h}_i = \phi(x_i) \in \mathbb{R}^D$, la representación de la bolsa se calcula como:

$$\boldsymbol{z} = \sum_{i=1}^{N} a_i \, \boldsymbol{h}_i \tag{2.1}$$

donde los pesos de atención a_i se obtienen mediante una red de atención con dos ramas paramétricas:

$$a_i = \frac{\exp\{\boldsymbol{w}^\top \left[\tanh(\boldsymbol{V}\boldsymbol{h}_i^\top) \odot \sigma(\boldsymbol{U}\boldsymbol{h}_i^\top)\right]\}}{\sum_{j=1}^{N} \exp\{\boldsymbol{w}^\top \left[\tanh(\boldsymbol{V}\boldsymbol{h}_j^\top) \odot \sigma(\boldsymbol{U}\boldsymbol{h}_j^\top)\right]\}} \tag{2.2}$$

La función tanh captura la importancia absoluta de cada instancia, mientras que la puerta sigmoide σ suprime las instancias que no contribuyen a la decisión. Este mecanismo se denomina *gated attention* y mejora la selectividad frente a la atención simple (sin puerta).

Los pesos $a_i \in [0, 1]$ son directamente interpretables como señal de explicabilidad: valores altos indican parches relevantes para la predicción.

2.4.3. CLAM-SB

Lu et al. [4] propusieron CLAM (*Clustering-constrained Attention MIL*), una extensión de ABMIL que introduce una pérdida auxiliar de clustering a nivel de instancia para guiar el aprendizaje de la atención de forma más discriminativa.

En la variante de una sola rama (CLAM-SB), el modelo selecciona los k parches con mayor atención como instancias pseudo-positivas y los k con menor atención como pseudo-negativas. Un clasificador binario por clase predice si cada instancia pertenece a su grupo correspondiente:

$$\mathcal{L}_{\text{inst}} = \frac{1}{2k} \sum_{i \in \text{top-}k \cup \text{bot-}k} \text{BCE}(\hat{y}_i, \tilde{y}_i) \tag{2.3}$$

La pérdida total combina la pérdida de clasificación de bolsa y la pérdida de instancia:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{bag}} + \lambda\, \mathcal{L}_{\text{inst}} \tag{2.4}$$

con $\lambda = 0{,}3$ en la implementación original. Esta pérdida auxiliar fuerza al mecanismo de atención a aprender representaciones más compactas y discriminativas desde las primeras épocas, lo que se traduce en mayor eficiencia de datos (*data efficiency*).

2.4.4. TransMIL

Shao et al. [17] propusieron TransMIL, que sustituye la agregación por atención independiente de ABMIL por un mecanismo de *self-attention* entre parches basado en transformers.

A diferencia de ABMIL, TransMIL modela las correlaciones espaciales y morfológicas entre instancias, lo que en principio permite capturar patrones de más alto nivel en la biopsia.

La arquitectura incorpora dos componentes novedosos:

- **Nyström Attention:** aproximación $O(N \cdot m)$ de la self-attention cuadrática $O(N^2)$, que selecciona m *landmarks* mediante submuestreo uniforme de las claves y valores, calculando la atención aproximada via pseudoinversa iterativa. Esto hace el modelo escalable a bolsas con miles de parches.
- **PPEG (*Pyramid Position Encoding Generator*):** codificación posicional 2D mediante convoluciones depthwise a tres escalas (3×3, 5×5, 7×7) aplicadas sobre el grid $\sqrt{N} \times \sqrt{N}$ de parches, añadiendo información de posición relativa entre instancias.

La representación de la bolsa se extrae del *token* CLS, que agrega información global a través de las capas TransLayer. La atención del CLS hacia los parches en la última capa es la señal de explicabilidad equivalente a los attention weights de ABMIL.

2.4.5. TransformerMIL

Frente a las arquitecturas MIL basadas en agregación por atención independiente (ABMIL) o en mecanismos Nyström de aproximación (TransMIL [17]), en este trabajo se propone una arquitectura denominada **TransformerMIL** que combina un *Transformer Encoder* estándar [3] con un mecanismo de *pooling* basado en *token* CLS.

La arquitectura proyecta los *embeddings* de entrada a un espacio oculto de menor dimensión, añade un *token* CLS aprendible, y los procesa con L capas de *multi-head self-attention*. La representación del CLS en la capa final se pasa por un clasificador MLP. La atención del CLS hacia los *tokens* de parche en la última capa de atención constituye la señal de explicabilidad de nivel 1. En comparación con ABMIL, TransformerMIL captura interacciones entre parches (no solo pesa cada parche de forma independiente), lo que en principio le permite modelar patrones de distribución espacial de las lesiones.

2.5. Explicabilidad en modelos de visión (XAI)

La explicabilidad (*Explainable AI*, XAI) es un requisito fundamental en aplicaciones médicas: no basta con que el modelo prediga correctamente, sino que debe ser posible justificar la predicción ante un clínico. En el contexto de las WSI, la XAI opera a dos niveles complementarios.

2.5.1. Attention maps

En los modelos basados en atención (ABMIL, CLAM, TransMIL), los pesos de atención a_i asignados a cada parche son directamente interpretables: parches con mayor peso han contribuido más a la predicción de la bolsa. Proyectando estos pesos sobre las coordenadas espaciales de los parches en la WSI se obtiene un mapa de calor (*heatmap*) que visualiza qué regiones anatómicas el modelo considera más informativas.

Este tipo de explicabilidad es intrínseca al modelo, a diferencia de los métodos post-hoc, lo que la hace computacionalmente eficiente y teóricamente coherente con el proceso de decisión. Sin embargo, tiene la limitación de que opera a resolución de parche (por ejemplo, 256×256 píxeles), sin indicar qué estructuras *dentro* del parche son relevantes.

2.5.2. Grad-CAM

Grad-CAM (*Gradient-weighted Class Activation Mapping*) [18] es un método post-hoc de explicabilidad que genera mapas de activación a nivel de píxel dentro de un parche, indicando qué regiones han activado la red neuronal para una clase dada.

Dado un parche x_i y una clase objetivo c, Grad-CAM calcula el gradiente de la puntuación de clase y^c con respecto a los mapas de activación A^k de la última capa convolucional:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2.5}$$

El mapa de calor Grad-CAM se obtiene como combinación lineal ponderada de los mapas de activación, seguida de una activación ReLU:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c \, A^k \right) \tag{2.6}$$

La ReLU elimina las activaciones con influencia negativa sobre la clase, conservando solo las regiones que incrementan la puntuación de c. El mapa resultante se interpola bilinealmente al tamaño del parche original para su visualización.

Combinado con los attention maps, Grad-CAM permite una cadena de explicabilidad completa: *qué parches* de la WSI importan (nivel de bolsa, vía atención) y *qué estructuras* dentro de esos parches activan la CNN (nivel de instancia, vía Grad-CAM).

2.5.3. ViT attention maps

Cuando el extractor de características es un Vision Transformer, existe un nivel adicional de explicabilidad: la atención interna del propio ViT dentro de cada parche. Al igual que en la arquitectura MIL, el ViT contiene un *token* CLS que agrega la información del parche completo mediante mecanismos de self-attention. Los pesos de atención del CLS hacia los 256 *tokens* de parche (para una imagen 224×224 con parches 14×14) en la última capa del ViT pueden interpretarse como un mapa de relevancia a nivel de píxel dentro del parche.

En el caso de H-optimus-0, la secuencia de *tokens* comprende el *token* CLS, seguido de 4 *register tokens* y posteriormente los 256 *tokens* de parche, totalizando 261 *tokens*. Los *register tokens* se sitúan entre el CLS y los *tokens* de parche, por lo que la extracción correcta de la atención requiere saltarlos y acceder directamente a los 256 *tokens* de parche.

Este nivel de explicabilidad, denominado Nivel 2B en este trabajo, complementa los mapas de atención de nivel de bolsa (Nivel 1) y la visualización de parches (Nivel 2A), completando una cadena de tres niveles de interpretabilidad desde la WSI completa hasta la estructura histológica intraparche.

2.6. Trabajos previos sobre el dataset KPMP

El *Kidney Precision Medicine Project* (KPMP) es una iniciativa del National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) de Estados Unidos, con el objetivo de caracterizar molecularmente las enfermedades renales mediante biopsia de referencia en pacientes con CKD y AKI [1]. El proyecto ha generado una cohorte de biopsias renales con datos multimodales (histología, transcriptómica espacial, proteómica, datos clínicos) que constituye uno de los recursos más completos del campo.

Desde el punto de vista del aprendizaje computacional, el trabajo más relevante es el publicado en *Scientific Reports* en 2025 [19], que aplica MIL con modelos fundacionales de histopatología sobre WSIs del KPMP. Los autores evalúan siete extractores de características (ResNet50, UNI, UNI2-h, Phikon, Prov-GigaPath, Virchow y Virchow2) con cuatro arquitecturas MIL (max pooling, ABMIL, TransMIL y CLAM), reportando un AUROC interno superior a 0,98 y externo superior a 0,94 con los mejores modelos. Sin embargo, su tarea de clasificación comprende únicamente tres clases: *Healthy controls*, *Acute Interstitial Nephritis* (AIN) y *Diabetic Kidney Disease* (DKD), usando exclusivamente WSIs teñidas con H&E.

Contribución original: hasta donde se conoce, el presente trabajo es el primero en abordar la clasificación simultánea de cuatro categorías diagnósticas — CKD, AKI, DM-R y *Healthy Reference* — sobre WSIs del KPMP con tinciones PAS y tricrómico. Esta combinación de clases es clínicamente más exigente que cualquier trabajo previo: AKI y CKD son entidades distintas que el estudio de referencia [19] no distingue, y la diferenciación DM-R/CKD avanzada representa un reto histológico reconocido incluso para patólogos expertos.

Otros trabajos destacados en el análisis computacional de biopsias renales incluyen métodos de segmentación de estructuras glomerulares mediante U-Net y variantes [13], si bien estas aproximaciones requieren anotaciones a nivel de píxel que no están disponibles de forma abierta en el KPMP, lo que motivó la elección de MIL en este trabajo.

Capítulo 3 Dataset

3.1. Kidney Precision Medicine Project (KPMP)

El *Kidney Precision Medicine Project* (KPMP) es una iniciativa del *National Institute of Diabetes and Digestive and Kidney Diseases* (NIDDK) de los Institutos Nacionales de Salud de Estados Unidos, cuyo objetivo es caracterizar molecularmente las enfermedades renales mediante biopsia en pacientes con CKD y AKI [1]. El proyecto integra datos multimodales de altísima resolución —histología, transcriptómica espacial, proteómica, metabolómica y datos clínicos— construyendo uno de los atlas de referencia de tejido renal humano más completos disponibles actualmente [1].

Una parte de los datos del KPMP es de acceso abierto (*Open Access*) a través del portal `atlas.kpmp.org`, y no requiere la firma de un acuerdo de uso de datos (*Data Use Agreement*, DUA). Las imágenes de biopsias digitalizadas se distribuyen en formato OME-TIFF, compatible con la librería OpenSlide [12] para su lectura en entornos de investigación.

Para la construcción del dataset de este trabajo se seleccionaron únicamente las biopsias de acceso abierto con tinciones PAS (*Periodic Acid-Schiff*) y Trichrome (tricrómico de Masson), por ser las tinciones más informativas para la evaluación de fibrosis intersticial, daño glomerular y alteraciones tubulares. El proceso de descarga y validación de las imágenes se realizó mediante scripts automatizados, identificando y descartando las imágenes no descargables o corruptas.

El dataset final comprende **710 biopsias renales** de alta resolución procedentes de biopsias de referencia (*reference biopsies*) del KPMP, correspondientes a pacientes con cuatro diagnósticos distintos.

3.2. Clases de diagnóstico

Las biopsias se clasifican en cuatro categorías diagnósticas, codificadas como etiquetas numéricas para el entrenamiento:

- **CKD (Chronic Kidney Disease, clase 0)**: Enfermedad renal crónica, caracterizada histológicamente por fibrosis intersticial y atrofia tubular (IFTA), esclerosis glomerular progresiva e infiltrado inflamatorio crónico [20]. Es la categoría con mayor número de biopsias en el dataset.

- **AKI (Acute Kidney Injury, clase 1)**: Lesión renal aguda, que se manifiesta histológicamente con daño tubular agudo (necrosis tubular aguda, NTA): células tubulares con núcleos aumentados, pérdida de la membrana basal tubular, detritos celulares intraluminales y, en algunos casos, cilindros hemáticos [21].
- **DM-R (Diabetes Mellitus-Related, clase 2)**: Nefropatía diabética, con hallazgos histológicos característicos como el engrosamiento de la membrana basal glomerular, la expansión mesangial difusa y, en estadios avanzados, los nódulos de Kimmelstiel-Wilson (depósitos nodulares de material amorfo eosinófilo en el mesangio) [6].
- **Healthy Reference (clase 3)**: Tejido renal sano de referencia, procedente de biopsias de donantes o voluntarios sin patología renal conocida. Representa el patrón histológico normal: glomérulos con arquitectura preservada, túbulos regulares y ausencia de fibrosis o inflamación.

La clasificación en cuatro categorías refleja las principales entidades clínicas para las que el KPMP ha reclutado pacientes, y constituye un problema de clasificación multiclase no trivial dado el solapamiento de algunas características histológicas entre las clases patológicas.

3.3. Tinciones histológicas: PAS y Trichrome

Las biopsias del dataset incluyen dos tipos de tinción principales:

PAS (Periodic Acid-Schiff): Tinción que reacciona con los polisacáridos de las membranas basales y el mesangio glomerular, tiñéndolos de magenta-rosa intenso sobre un fondo azul pálido. Permite visualizar con claridad la arquitectura glomerular, las membranas basales tubulares y la expansión mesangial, siendo especialmente útil para el diagnóstico de nefropatías glomerulares como la DM-R.

Trichrome (tricrómico de Masson): Tinción tricromial que diferencia el tejido conectivo (colágeno) en azul o verde, el citoplasma en rojo y los núcleos en negro. Es el estándar de referencia para la evaluación de la fibrosis intersticial y la atrofia tubular (IFTA), hallazgo clave en CKD. La intensidad del azul intersticial permite una cuantificación visual semi-cuantitativa del grado de fibrosis.

El uso combinado de ambas tinciones en el dataset aporta información complementaria: PAS para los cambios glomerulares y DM-R, Trichrome para la fibrosis en CKD. El modelo recibe esta diversidad de tinciones durante el entrenamiento sin distinción explícita del tipo de tinción, aprendiendo representaciones invariantes a la misma.

3.4. Distribución del dataset y desbalance de clases

La distribución de las 710 biopsias entre las cuatro clases se muestra en la Tabla 1 y la Figura 1.

Tabla 1: Distribución de biopsias por clase diagnóstica en el dataset KPMP utilizado.

Clase	**N slides**	**%**
CKD (clase 0)	∼220	∼31 %
AKI (clase 1)	∼220	∼31 %
DM-R (clase 2)	∼170	∼24 %
Healthy (clase 3)	∼100	∼14 %
Total	**710**	**100 %**

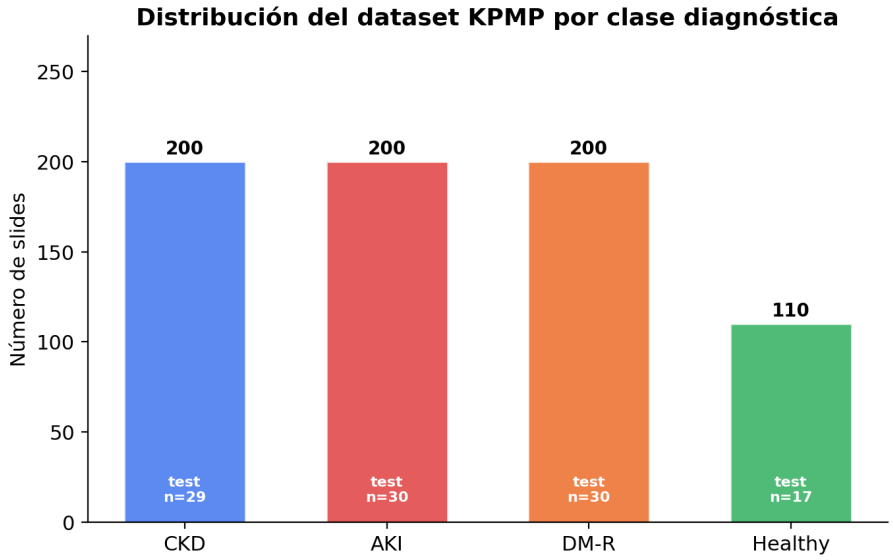


Figura 1: Distribución de biopsias por clase diagnóstica en el dataset KPMP utilizado. Se indica entre paréntesis el número de slides en el conjunto de test para cada clase.

El dataset presenta un desbalance moderado: la clase *Healthy Reference* tiene aproximadamente la mitad de representantes que CKD y AKI, mientras que DM-R ocupa una posición intermedia. Este desbalance se mitiga durante el entrenamiento mediante el uso de pérdida de entropía cruzada ponderada por clase, donde el peso de cada clase es inversamente proporcional a su frecuencia (véase la Sección 4.5.1).

Cabe destacar que un pequeño número de biopsias (∼9-10 slides) presentaron problemas durante la descarga o la extracción de parches y fueron descartadas del dataset efectivo, quedando el total final en 710 slides procesables.

3.5. División train / validación / test

El dataset se divide en tres particiones de forma estratificada por clase para garantizar que la distribución de clases se mantiene en cada partición. La división utilizada en los experimentos finales es la siguiente:

Tabla 2: Distribución de biopsias por partición del dataset.

Partición	**N slides**	**% total**	**Uso**
Train	545	76,8 %	Entrenamiento del modelo
Validación	59	8,3 %	Selección del mejor checkpoint
Test	106	14,9 %	Evaluación final del sistema
Total	**710**	**100 %**	

El conjunto de test (106 slides) se reserva estrictamente para la evaluación final y nunca se utiliza en ninguna decisión de diseño, selección de hiperparámetros ni de checkpoint. La distribución de clases en el test se refleja en la Tabla 3:

Tabla 3: Distribución del conjunto de test por clase.

Clase	**N slides (test)**
CKD (clase 0)	29
AKI (clase 1)	30
DM-R (clase 2)	30
Healthy (clase 3)	17
Total	**106**

La partición de validación se utiliza para el criterio de *early stopping*: el entrenamiento se detiene si el F1-macro sobre validación no mejora durante 30 épocas consecutivas, y el checkpoint con mayor F1 de validación es el que se carga para la evaluación final sobre test.

Capítulo 4 Metodología

4.1. Visión general del pipeline

El sistema desarrollado en este trabajo sigue un pipeline de dos etapas bien diferenciadas, ilustrado de forma esquemática a continuación:

1. **Etapa offline (preprocesamiento y extracción de *embeddings*):** Cada WSI se somete a tiling con filtrado de fondo, y los parches de tejido resultantes se procesan con el *backbone* de fundamento para obtener un vector de *embedding* por parche. Los *embeddings* se almacenan en disco en formato NumPy (`.npy`) como un array de forma $N \times D$, donde N es el número de parches válidos de la slide y D es la dimensión del *embedding* (1.536 para H-optimus).

2. **Etapa online (entrenamiento e inferencia del modelo MIL):** Durante el entrenamiento, el modelo MIL carga en cada iteración el array de *embeddings* de una slide, submuestrea hasta 256 parches aleatoriamente si $N > 256$, y realiza la predicción de clase junto con el cálculo de la pérdida. En inferencia, se aplica TTA mediante K pases con diferentes submuestras.

Esta separación de etapas es fundamental para hacer viable el entrenamiento en Google Colab Pro: el paso más costoso computacionalmente (extracción de *embeddings* con H-optimus, un modelo de 1.100M parámetros) se realiza una sola vez, y el entrenamiento del modelo MIL, mucho más ligero, puede iterarse con distintas configuraciones.

4.2. Preprocesamiento de WSI

4.2.1. Tiling con OpenSlide

Cada WSI se divide en parches (*patches*) de tamaño fijo de 256×256 píxeles extraídos al nivel de mayor resolución disponible mediante la librería OpenSlide [12]. Se utiliza una ventana deslizante sin solapamiento, con desplazamiento igual al tamaño del parche, lo que garantiza la cobertura completa del tejido sin redundancia.

El número de parches extraídos varía significativamente entre slides en función del tamaño del tejido: oscila entre pocos centenares y varios miles de parches por slide. Durante el entrenamiento se utiliza submuestreo aleatorio a 256 parches como máximo, lo que a su vez hace posible el TTA.

4.2.2. Filtrado de fondo: HSV y varianza del Laplaciano

Un desafío habitual en el análisis de WSI es que una fracción significativa de los parches extraídos corresponde a regiones de fondo (cristal vacío, grasa, artefactos) que no contienen tejido informativo. Incluir estos parches en el *bag* puede diluir las representaciones y perjudicar el rendimiento del modelo.

Para mitigar este problema se aplica un filtrado de dos etapas:

Etapa 1 — Filtrado HSV: Se convierte el parche al espacio de color HSV y se aplica una máscara de tejido basada en el canal de saturación: se descartan los parches en los que más del 70 % de los píxeles tienen saturación baja (fondo blanco) o matiz homogéneo (artefactos de una sola tonalidad). Este filtro retiene los parches con contenido cromático variado, típicos del tejido teñido.

Etapa 2 — Varianza del Laplaciano: El filtro laplaciano es un operador diferencial de segundo orden que aproxima la segunda derivada de la intensidad de la imagen. Para cada píxel, se aplica la convolución con el núcleo discreto estándar

$$K = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

que pondera la diferencia entre la intensidad del píxel y la media de sus cuatro vecinos. En regiones de fondo o tejido desenfocado, donde la intensidad varía suavemente, las respuestas del filtro son pequeñas; en regiones con bordes y texturas nítidas, las respuestas son grandes. La varianza de estas respuestas sobre todos los píxeles del parche resume en un único escalar el grado de nitidez global del parche. Esta métrica no tiene cota superior fija: en imágenes histológicas de 8 bits, los parches de fondo uniforme presentan valores próximos a 0, los parches borrosos o de tejido escaso se sitúan generalmente entre 10 y 50, y los parches de tejido bien enfocado suelen superar los 100, pudiendo alcanzar varios centenares en regiones de alta heterogeneidad morfológica. Los parches con varianza por debajo de un umbral de 45 se descartan por corresponder a regiones borrosas o de baja textura. Este umbral se determinó empíricamente mediante un análisis visual de los parches rechazados y aceptados [22].

La combinación de ambos filtros es necesaria porque algunos parches de fondo pasan el filtro HSV (por tener ligeras variaciones de color) pero no el Laplaciano (al ser lisos), y viceversa (algunos parches de tejido legítimo tienen coloración uniforme). El filtrado híbrido recuperó aproximadamente 32 slides que habrían sido descartadas al aplicar únicamente el

filtro HSV.

4.2.3. Caché de coordenadas

Para evitar repetir el proceso de extracción de coordenadas en cada ejecución, las coordenadas (x, y) de cada parche válido se almacenan en un archivo de coordenadas junto a la WSI correspondiente. Este archivo actúa como caché: si ya existe, las coordenadas se cargan directamente sin releer la WSI.

4.3. Modelos fundacionales como extractores de características

Una vez determinados los parches válidos de cada slide, se extraen los *embeddings* mediante el modelo fundacional seleccionado. En este trabajo se evalúan cuatro extractores:

- **ResNet50** [7]: Red residual de 50 capas preentrenada en ImageNet-1k. Produce *embeddings* de 2.048 dimensiones. Se usa como *baseline* clásico.
- **CONCH** [8]: ViT-B/16 visión-lenguaje, 512-dim. Acceso mediante la API de Hugging Face.
- **Phikon** [10]: ViT-B/16 de Owkin, preentrenado con MIM sobre 40M parches histológicos, 768-dim.
- **H-optimus-0** [2]: ViT-G/14 de Bioptimus, 1.100M parámetros, 1.536-dim. Backbone seleccionado para el sistema final.

En todos los casos, el *backbone* se usa en modo de inferencia pura, sin cálculo de gradientes, con los pesos congelados. La extracción se realiza parche a parche o en mini-batches con tamaño de batch 32, procesando todos los parches válidos de cada slide y guardando el array de *embeddings* resultante.

Para H-optimus-0, el modelo se carga desde Hugging Face Hub con sus pesos preentrenados en modo de evaluación.

4.4. Arquitecturas MIL implementadas

4.4.1. ABMIL

La arquitectura ABMIL (*Attention-Based Multiple Instance Learning*) [11] implementada consiste en:

1. Una capa de proyección lineal $\phi : \mathbb{R}^D \to \mathbb{R}^{256}$ con activación ReLU y *dropout*.
2. Un mecanismo de *gated attention* (Ecuación 2.2) que produce pesos $a_i \in [0, 1]$ para cada parche.
3. Un clasificador lineal $\mathbb{R}^{256} \to \mathbb{R}^4$ sobre la representación de la bolsa $\boldsymbol{z} = \sum_i a_i \boldsymbol{h}_i$.

4.4.2. TransformerMIL

La arquitectura TransformerMIL propuesta en este trabajo extiende el paradigma MIL con un mecanismo de *self-attention* global entre todos los parches de la bolsa. Sus componentes son:

1. **Proyección de entrada**: Capa lineal $\mathbb{R}^D \to \mathbb{R}^{d_{\text{hidden}}}$ (con $d_{\text{hidden}} = 256$) aplicada a cada *embedding* de parche.
2. **Token CLS aprendible**: Vector $\boldsymbol{c} \in \mathbb{R}^{d_{\text{hidden}}}$ inicializado aleatoriamente y prepuesto a la secuencia de parches proyectados.
3. **Transformer Encoder**: $L = 2$ capas de codificador Transformer estándar [3] con $H = 4$ cabezas de atención, dimensión intermedia $d_{\text{ff}} = 512$ y tasa de *dropout* de 0,4.
4. **Clasificador**: Capa lineal $\mathbb{R}^{d_{\text{hidden}}} \to \mathbb{R}^4$ sobre la representación del *token* CLS en la última capa.

La configuración final del sistema SOTA emplea dimensión de entrada 1.536 (H-optimus-0), dimensión oculta 256, 2 capas Transformer, 4 cabezas de atención y tasa de *dropout* de 0,4.

La señal de explicabilidad de TransformerMIL se obtiene a partir de la similitud coseno entre la representación del CLS y cada *token* de parche tras el encoder:

$$a_i = \frac{\boldsymbol{c}_{\text{final}} \cdot \boldsymbol{h}_i^{\text{final}}}{\|\boldsymbol{c}_{\text{final}}\| \cdot \|\boldsymbol{h}_i^{\text{final}}\|} \tag{4.1}$$

Este valor es directamente interpretable como la relevancia del parche i para la decisión de clasificación.

4.4.3. CLAM-SB

CLAM-SB (*Clustering-constrained Attention Multiple Instance Learning*, variante single-branch) [23] extiende ABMIL añadiendo una pérdida auxiliar a nivel de instancia que supervisa qué parches individuales son más discriminativos, sin requerir anotaciones de instancia. Sus componentes son:

1. **Compresión de características**: Capa lineal $\mathbb{R}^D \to \mathbb{R}^{512}$ con ReLU y *dropout*, que reduce la dimensión del *embedding* antes de la atención.

2. **Gated attention**: Mecanismo idéntico al de ABMIL (Ecuación 2.2) con dimensión de atención 256, que produce pesos $a_i \in [0, 1]$ para cada parche y la representación de bolsa $\boldsymbol{z} = \sum_i a_i \boldsymbol{h}_i$.

3. **Clasificador de bolsa**: Capa lineal $\mathbb{R}^{512} \to \mathbb{R}^4$ sobre $\boldsymbol{z}$, cuya pérdida $\mathcal{L}_{\text{bag}}$ es la entropía cruzada ponderada estándar.

4. **Pérdida de instancia (clustering)**: Para la clase predicha $\hat{c}$, se seleccionan los $k = 8$ parches con mayor atención como pseudo-positivos y los $k = 8$ con menor atención como pseudo-negativos. Sobre cada parche seleccionado se aplica un clasificador binario $f_{\hat{c}} : \mathbb{R}^{512} \to \{0, 1\}$ entrenado con estas pseudo-etiquetas:

$$\mathcal{L}_{\text{inst}} = \frac{1}{2k} \left(\sum_{i \in \text{top-}k} \text{CE}(f_{\hat{c}}(\boldsymbol{h}_i),\, 1) + \sum_{i \in \text{bottom-}k} \text{CE}(f_{\hat{c}}(\boldsymbol{h}_i),\, 0) \right) \tag{4.2}$$

La pérdida total combina ambos términos con peso $\lambda = 0{,}3$:

$$\mathcal{L}_{\text{CLAM}} = \mathcal{L}_{\text{bag}} + \lambda\, \mathcal{L}_{\text{inst}} \tag{4.3}$$

La intuición es que la señal de clustering fuerza al modelo a concentrar la atención en parches con características discriminativas, mejorando la selectividad de la atención en datasets grandes.

4.4.4. TransMIL

TransMIL [17] es una arquitectura Transformer para MIL que incorpora codificación de posición 2D mediante un módulo piramidal de convoluciones (*Pyramid Position Encoding*

Generator, PPEG), preservando así la estructura espacial de los parches dentro de la bolsa. Sus componentes son:

1. **Proyección de entrada**: Capa lineal $\mathbb{R}^D \to \mathbb{R}^d$ (con $d = 512$) aplicada independientemente a cada *embedding* de parche.

2. **PPEG (codificación posicional 2D)**: Módulo que inyecta información de posición relativa mediante tres ramas de convoluciones depthwise en paralelo con kernels 7×7, 5×5 y 3×3, sumadas a la representación proyectada. Esto permite al modelo razonar sobre la disposición espacial de los parches sin coordenadas explícitas.

3. **Token CLS aprendible**: Vector $\boldsymbol{c} \in \mathbb{R}^d$ prepuesto a la secuencia, que agrega la información global de la bolsa a través del mecanismo de atención.

4. **Transformer Encoder**: $L = 2$ capas *Pre-LN* [24] con $H = 8$ cabezas de atención, dimensión intermedia $d_{\text{ff}} = 2048$ y *dropout*=0,1. La variante Pre-LN estabiliza el entrenamiento al aplicar la normalización de capa antes de cada sublayer, evitando los problemas de gradientes que presenta la implementación original con NystromAttention en *embeddings* de alta dimensión.

5. **Clasificador**: Capa lineal $\mathbb{R}^d \to \mathbb{R}^4$ sobre la representación del *token* CLS en la última capa del encoder.

La señal de explicabilidad se obtiene, al igual que en TransformerMIL, como la similitud coseno entre el *token* CLS final y cada *token* de parche (Ecuación 4.1).

4.5. Estrategia de entrenamiento

4.5.1. Función de pérdida ponderada

Para compensar el desbalance de clases se emplea entropía cruzada ponderada:

$$\mathcal{L} = -\sum_{k=0}^{3} w_k \, y_k \log \hat{p}_k \tag{4.4}$$

donde $w_k \propto 1/N_k$ es el peso de la clase k (inversamente proporcional al número de muestras N_k de esa clase en el conjunto de entrenamiento), e $\hat{p}_k$ es la probabilidad predicha para la clase k.

4.5.2. Optimizador y scheduler

El entrenamiento utiliza el optimizador AdamW [25] con tasa de aprendizaje inicial $\text{lr} = 10^{-4}$ y decaimiento de pesos $\lambda = 10^{-4}$.

El scheduler de tasa de aprendizaje es *cosine annealing* [26]: la tasa de aprendizaje decrece siguiendo una coseno desde $\text{lr}_{\max}$ hasta $\text{lr}_{\min} = 10^{-6}$ a lo largo de $T_{\max}$ épocas, sin reinicios (*warm restarts* desactivados).

4.5.3. Early stopping

El entrenamiento se ejecuta durante un máximo de 100 épocas con early stopping de paciencia $p = 30$: si el F1-macro en el conjunto de validación no mejora durante 30 épocas consecutivas, el entrenamiento se detiene y se carga el *checkpoint* con la mejor métrica de validación. Esta estrategia evita el sobreajuste mientras permite al modelo suficiente tiempo para converger en valles secundarios.

4.5.4. Submuestreo de parches

Dado que el número de parches por slide varía ($N \in [50, 2000+]$), durante el entrenamiento se submuestrea aleatoriamente un máximo de 256 parches por slide en cada época. Este submuestreo actúa como una forma implícita de augmentación de datos, ya que la misma slide se ve representada por subconjuntos distintos de parches en diferentes épocas.

4.6. Test-Time Augmentation (TTA)

El TTA es una técnica de inferencia que consiste en realizar K pasadas hacia adelante del modelo sobre la misma muestra con variaciones aleatorias, y promediar las predicciones resultantes [27]. En el contexto de este trabajo, la variación consiste en submuestrar K subconjuntos distintos de 256 parches de los N disponibles en la slide:

1. Para cada slide de test con N parches, se realizan K submuestreos aleatorios sin reemplazo de 256 parches.

2. Para cada submuestreo, se obtiene el vector de probabilidades $\boldsymbol{p}^{(k)} \in \mathbb{R}^4$ mediante una pasada del modelo.

3. La predicción final se obtiene como el promedio de los K vectores de probabilidad:

$$\bar{\boldsymbol{p}} = \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{p}^{(k)} \tag{4.5}$$

4. La clase predicha es $\hat{y} = \arg\max_c \bar{p}_c$.

Para las slides con $N \leq 256$ parches, el TTA no tiene efecto (todos los submuestreos son idénticos), pero para el 60 % de las slides del conjunto de test que sí tienen más parches, el TTA estabiliza la predicción reduciendo la varianza debida a la aleatoriedad del submuestreo. En los experimentos, $K = 5$ demostró el mejor balance entre estabilización y coste computacional.

El TTA no utiliza en ningún momento las etiquetas del test, por lo que no constituye fuga de información (*data leakage*), y es una técnica estándar en sistemas de inferencia para diagnóstico asistido.

4.7. Módulos de explicabilidad (XAI)

4.7.1. Nivel 1 — Mapa de calor de atención sobre la WSI

Una vez obtenidas las puntuaciones de atención a_i (Ecuación 4.1) para cada parche de la slide, se proyectan de vuelta sobre las coordenadas espaciales originales para construir un mapa de calor a escala de la WSI:

1. Se normaliza la atención: $\tilde{a}_i = (a_i - \min_j a_j)/(\max_j a_j - \min_j a_j)$.
2. Se crea un array 2D del tamaño del thumbnail de la WSI (tipicamente 1.000×1.000 píxeles), inicializado a cero.
3. Para cada parche i, se rellena la región correspondiente en el array con $\tilde{a}_i$.
4. El array se colorea con un mapa de colores espectral (frío=baja atención, cálido=alta atención) y se superpone con transparencia sobre el thumbnail RGB de la WSI.

4.7.2. Nivel 2A — Visualización de los parches más relevantes

Los $K_{\text{top}} = 9$ parches con mayor atención se extraen de la WSI original a resolución completa y se visualizan en una cuadrícula RGB. Esta visualización permite inspeccionar

directamente el contenido histológico que el modelo considera más informativo para cada clase.

4.7.3. Nivel 2B — Atención interna del ViT (register-aware)

Para cada parche de alta atención, se extrae el mapa de atención interno del ViT *backbone* (H-optimus-0) interceptando la capa de atención de la última capa del ViT. El procedimiento es:

1. Se captura la matriz de atención $\boldsymbol{A} \in \mathbb{R}^{H \times N_{\text{tokens}} \times N_{\text{tokens}}}$ (con $H = 24$ cabezas y $N_{\text{tokens}} = 261$).
2. Se extrae la atención del CLS hacia los 256 *tokens* de parche, saltando los 4 *register tokens*, obteniendo un vector $\in \mathbb{R}^{24 \times 256}$.
3. Se promedia sobre las 24 cabezas, obteniendo un vector $\in \mathbb{R}^{256}$.
4. Se reordena en una cuadrícula 16×16 (para patches 14×14 con imagen 224×224) y se interpola al tamaño del parche original (256×256).
5. El mapa resultante se superpone sobre el parche RGB como un heatmap.

Capítulo 5 Experimentos y resultados

5.1. Configuración experimental

Todos los experimentos se ejecutan en Google Colab Pro con una GPU NVIDIA A100 (40 GB VRAM) o T4 (16 GB VRAM), según disponibilidad.

La semilla de aleatoriedad se fija a 42 para la división train/val/test y para la inicialización de los pesos del modelo MIL, garantizando la reproducibilidad. Los submuestreos de parches durante el entrenamiento no se fijan (son intrínsecos a la técnica de augmentación implícita). Los hiperparámetros comunes a todos los experimentos se recogen en la Tabla 4.

Tabla 4: Hiperparámetros comunes a todos los experimentos.

Hiperparámetro	**Valor**
max_patches	256
Optimizador	AdamW (lr $= 10^{-4}$, $\lambda = 10^{-4}$)
Scheduler	Cosine annealing ($T_{\max} = 100$, $\eta_{\min} = 10^{-6}$)
Early stopping (paciencia)	30 épocas
Máximo de épocas	100
Métrica de selección	F1-macro (validación)
Función de pérdida	Cross-entropy ponderada por clase

5.2. Experimento 1: Comparativa de backbones con ABMIL

En primer lugar se evalúa la influencia del *backbone* en el rendimiento del modelo ABMIL, manteniendo la misma arquitectura de agregación. Los resultados en el conjunto de test se recogen en la Tabla 5 y se visualizan en la Figura 2.

Tabla 5: Comparativa de *backbones* con arquitectura ABMIL en el test set (106 slides).

Backbone	**F1-macro**	**Accuracy**
ResNet50 (ImageNet)	≈0,30	≈0,35
CONCH	0,798	0,783
UNI	≈0,581	≈0,580
Phikon	0,905	0,896
H-optimus-0	0,884	0,877

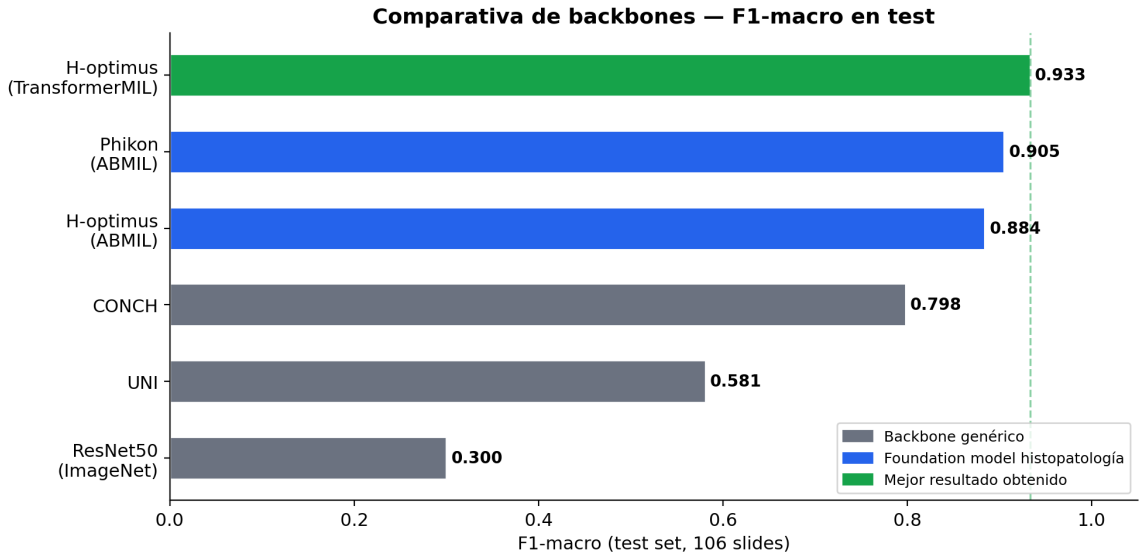


Figura 2: Comparativa de F1-macro en test para los distintos *backbones* evaluados con arquitectura ABMIL. Los modelos fundacionales de histopatología superan ampliamente a ResNet50 preentrenado en ImageNet.

Los resultados evidencian una diferencia abismal entre ResNet50 (entrenado en ImageNet) y los modelos fundacionales de histopatología. CONCH y UNI ofrecen mejoras sustanciales sobre ResNet50, pero se sitúan a distancia de Phikon y H-optimus-0, que han sido preentrenados con órdenes de magnitud más datos histológicos.

Sorprendentemente, con ABMIL Phikon (768 dimensiones) supera a H-optimus-0 (1.536 dimensiones). Esto sugiere que el mecanismo de agregación independiente de ABMIL no es capaz de aprovechar plenamente la mayor riqueza representacional de H-optimus-0, y que se requiere una arquitectura de agregación más potente.

5.3. Experimento 2: Comparativa de arquitecturas MIL con Phikon y H-optimus

Se comparan las dos arquitecturas MIL (ABMIL y TransformerMIL) con los dos mejores *backbones*. Los resultados se resumen en la Tabla 6 y la Figura 3.

Tabla 6: Comparativa de arquitecturas MIL con Phikon y H-optimus en el test set.

Backbone	Arquitectura MIL	TTA	F1-macro	Accuracy
Phikon	ABMIL	—	0,905	0,896
Phikon	ABMIL	$K = 5$	0,925	0,911
Phikon	TransformerMIL	—	0,905	0,896
H-optimus-0	ABMIL	—	0,884	0,877
H-optimus-0	TransformerMIL	$K = 5$	**0,933**	**0,925**

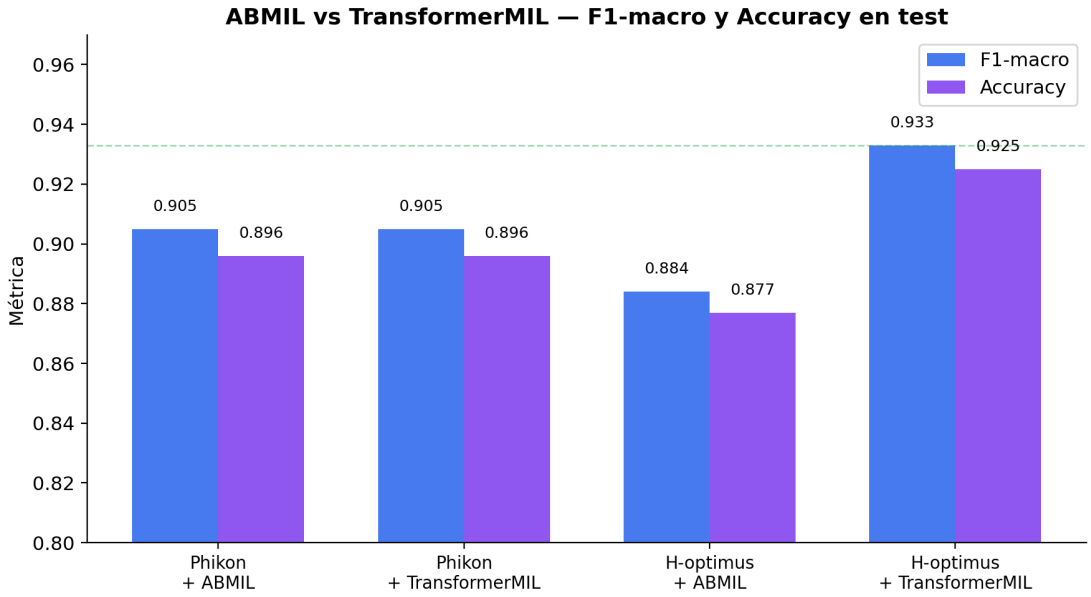


Figura 3: F1-macro y Accuracy en test para las combinaciones de *backbone* y arquitectura MIL evaluadas. TransformerMIL supera a ABMIL únicamente con H-optimus-0, cuya mayor dimensionalidad de *embedding* aprovecha el mecanismo de self-attention entre parches.

Los resultados muestran que sin TTA, ABMIL y TransformerMIL alcanzan el mismo rendimiento con Phikon (F1=0,905), lo que sugiere que la capacidad representacional del *backbone* es el cuello de botella en ese régimen. ABMIL con Phikon se beneficia especialmente del TTA ($K = 5$), subiendo a F1=0,925 (+2,0 puntos), lo que indica que la varianza debida al submuestreo aleatorio de parches es el principal factor limitante en ese modelo. Con H-optimus-0, TransformerMIL + TTA alcanza F1=0,933 (+4,9 puntos sobre ABMIL sin TTA), aprovechando la mayor riqueza semántica de los *embeddings* 1.536-dimensionales.

La combinación H-optimus-0 + TransformerMIL se convierte en el sistema de mayor rendimiento (**SOTA de este trabajo**), representando una mejora del **+47 %** en F1-macro

respecto al *baseline* ResNet50.

5.3.1. Arquitecturas MIL de referencia: CLAM-SB y TransMIL

Con el fin de contextualizar los resultados anteriores, se evaluaron adicionalmente dos arquitecturas MIL ampliamente citadas en la literatura: CLAM-SB [23] y TransMIL [17], ambas en modo *embedding* con los cuatro *backbones*. La Tabla 7 resume el mejor resultado de cada arquitectura.

Tabla 7: Comparativa de F1-macro en test (106 slides) entre las cuatro arquitecturas evaluadas.

Arquitectura	**Mejor F1-macro**	**Backbone**	**TTA**
ABMIL	0,925	Phikon	$K = 5$
CLAM-SB	0,847	Phikon	—
TransMIL (Shao 2021)	0,857	H-optimus	—
TransformerMIL (nuestro)	**0,933**	H-optimus	$K = 5$

CLAM-SB introduce una pérdida auxiliar de clustering a nivel de instancia que, en datasets grandes, mejora la discriminabilidad de la atención. Sin embargo, con $N \approx 550$ WSIs, esta señal adicional provoca sobreajuste severo: el modelo alcanza F1=1,0 en entrenamiento pero no generaliza, resultando inferior a ABMIL en tres de los cuatro *backbones* evaluados. TransMIL, por su parte, obtiene resultados competitivos con H-optimus (F1=0,857), pero no supera a ABMIL en ningún *backbone* y queda por debajo de TransformerMIL en todos los casos comparables, lo que confirma que el mecanismo de self-attention global con *token* CLS es más adecuado para esta tarea que la codificación posicional piramidal de TransMIL.

5.3.2. Curvas de entrenamiento: H-optimus + TransformerMIL

El entrenamiento del modelo SOTA duró 65 épocas, con el mejor *checkpoint* en la época 35 (F1 de validación = 0,921), detenido por el criterio de parada temprana con paciencia de 30 épocas. La Figura 4 muestra las curvas de pérdida de entrenamiento y F1 de validación a lo largo del entrenamiento.

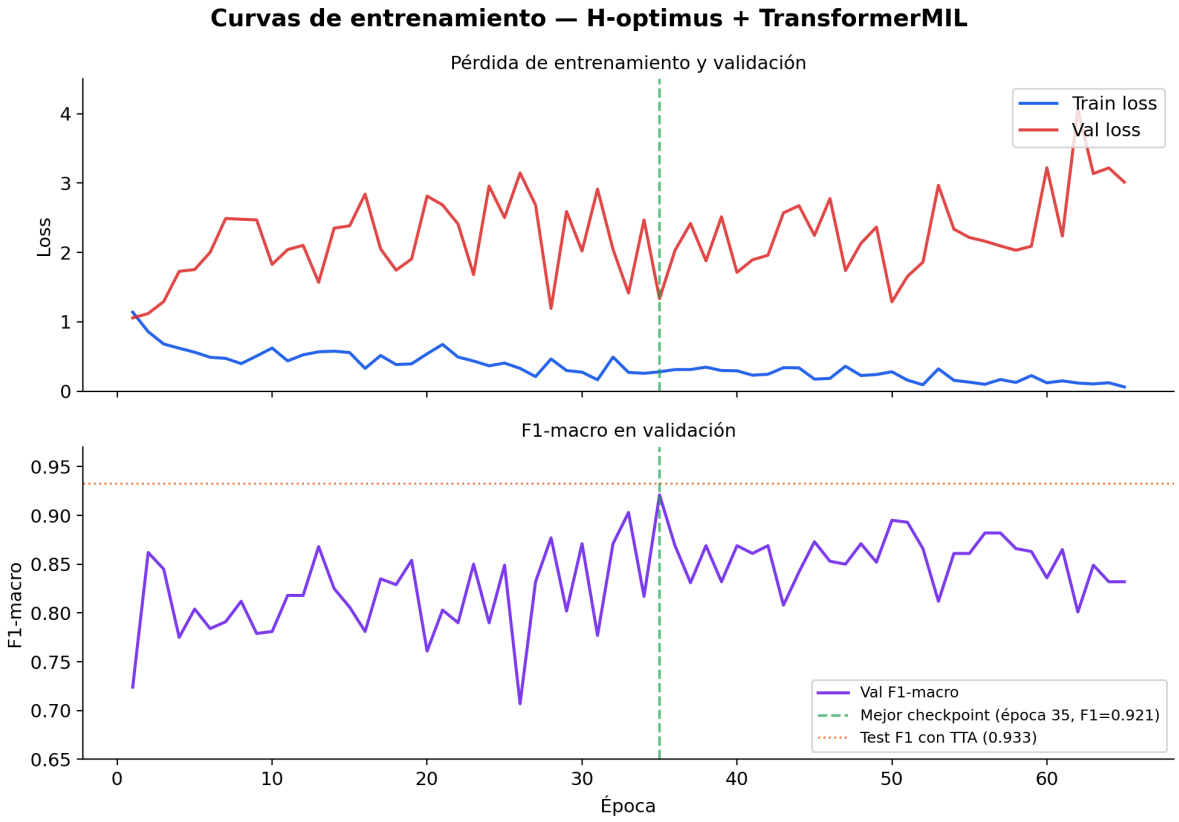


Figura 4: Curvas de entrenamiento del modelo H-optimus + TransformerMIL. La pérdida de entrenamiento decrece de 1,14 a 0,06, mientras el F1 de validación alcanza su máximo en la época 35 (0,921) y se mantiene ruidoso por el pequeño tamaño del conjunto de validación.

Se observa un patrón de sobreajuste (*overfitting*) moderado: la pérdida de entrenamiento converge hacia valores muy bajos (≈0,06) mientras que la pérdida de validación diverge hasta valores próximos a 4,0. Sin embargo, el F1 de validación se mantiene razonablemente alto y el F1 en test final (0,925 sin TTA, 0,933 con TTA) confirma que el modelo ha aprendido representaciones genuinamente discriminativas.

Un aspecto destacable es el **elevado nivel de ruido en la pérdida de entrenamiento**: en lugar de descender de forma suave y monotónica, la curva presenta oscilaciones pronunciadas, especialmente durante las primeras 40 épocas. Este comportamiento refleja la fragilidad inherente al modelo dados los recursos de datos disponibles: con únicamente 545 slides de entrenamiento y bolsas de 256 parches submuestreados aleatoriamente en cada época, cada minibatch es cualitativamente diferente del anterior, impidiendo una convergencia estable. Este patrón de ruido es un indicador directo de la necesidad de ampliar el dataset para al-

canzar un entrenamiento más robusto; con más slides, la varianza de cada minibatch sería menor y la curva de pérdida convergiría con mayor suavidad.

El ruido del *val_f1* es igualmente inherente al pequeño tamaño del conjunto de validación (59 slides), donde la variación de un solo caso supone ∼1,7 % de cambio en la métrica, amplificando el ruido estadístico en la señal de selección del *checkpoint*.

5.4. Experimento 3: Test-Time Augmentation

Se evalúa el efecto del TTA sobre el modelo H-optimus + TransformerMIL variando el número de pasadas K. Los resultados se muestran en la Tabla 8.

Tabla 8: Efecto del TTA sobre el F1-macro en test (K pasadas de submuestreo).

K	**F1-macro**	**Accuracy**
1 (sin TTA)	0,901–0,925*	0,896–0,925*
5	**0,933**	**0,925**
10	0,933	0,925
20	0,933	0,925

*El resultado sin TTA varía entre ejecuciones por la aleatoriedad del submuestreo.

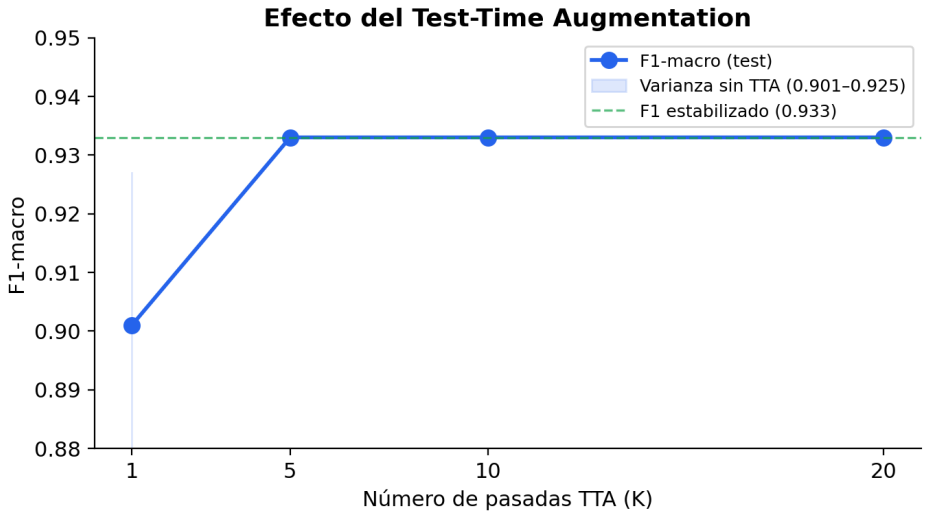


Figura 5: Efecto del número de pasadas TTA (K) sobre el F1-macro en test. Con $K = 5$ ya se alcanza el rendimiento máximo y se elimina la varianza debida al submuestreo aleatorio de parches.

Los resultados demuestran que el TTA tiene dos efectos positivos: (1) estabiliza la predicción, eliminando la varianza debida al submuestreo aleatorio, y (2) mejora el rendimiento en ∼0,008 F1 al recuperar la capacidad del modelo que se pierde cuando un único submuestreo omite parches críticos. Con $K = 5$ ya se alcanza el rendimiento máximo; aumentar K a 10 o 20 no produce mejora adicional, confirmando que el error de predicción con TTA está dominado por los casos inherentemente difíciles para el modelo, no por la varianza del submuestreo.

El 60 % de las slides del test tienen $N > 256$ parches (rango 257–2.000+), lo que hace el TTA efectivo para la mayoría de los casos.

5.5. Resultados del sistema final

El sistema final (H-optimus-0 + TransformerMIL + TTA $K = 5$) obtiene los siguientes resultados en el conjunto de test de 106 slides:

- **F1-macro: 0,933**
- **Accuracy: 0,925** (98/106 slides clasificadas correctamente)

El informe de clasificación detallado por clase se presenta en la Tabla 9 y la Figura 6.

Tabla 9: Informe de clasificación del sistema final (H-optimus + TransformerMIL + TTA $K = 5$).

Clase	Precision	Recall	F1-score	Support
CKD	0,87	0,90	0,88	29
AKI	0,96	0,90	0,93	30
DM-R	0,90	0,93	0,92	30
Healthy	1,00	1,00	1,00	17
Macro avg	**0,93**	**0,93**	**0,933**	106

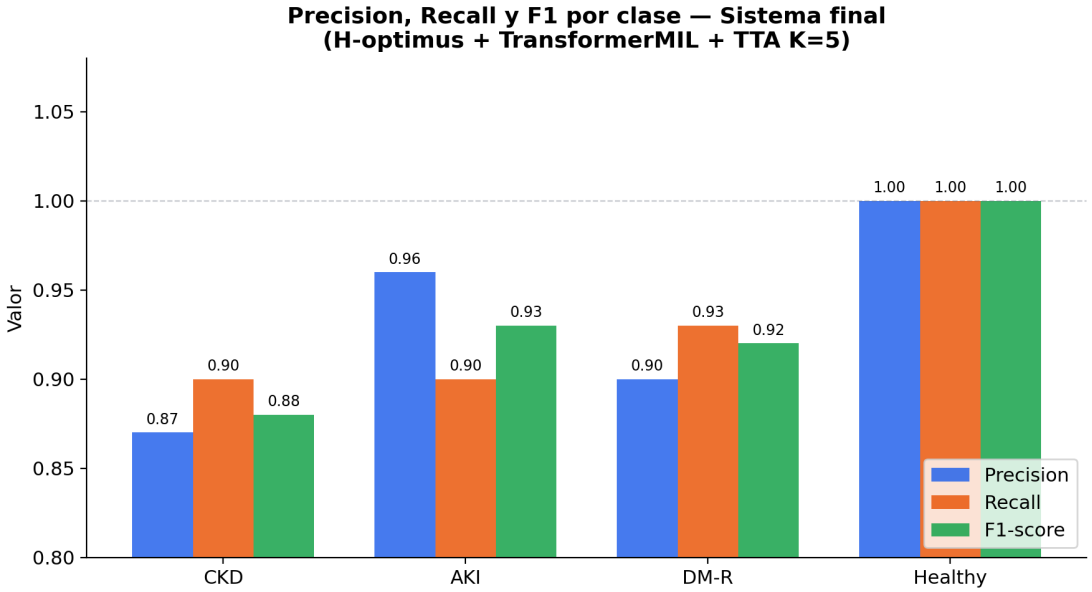


Figura 6: Precision, Recall y F1-score por clase diagnóstica del sistema final (H-optimus + TransformerMIL + TTA $K = 5$). La clase *Healthy* se clasifica con perfección absoluta; CKD presenta la mayor dificultad.

Los resultados por clase muestran un rendimiento heterogéneo: *Healthy Reference* se clasifica con perfección absoluta (F1=1,00), mientras que CKD presenta el mayor número de errores (F1=0,88). La clase más difícil de distinguir es CKD, cuyo patrón histológico (fibrosis intersticial progresiva) puede solaparse con estadios avanzados de DM-R. AKI y DM-R obtienen resultados intermedios con alta precisión.

La Figura 7 y la Figura 8 muestran las matrices de confusión para los dos modelos basados en H-optimus-0, reconstruidas a partir del informe de clasificación por clase. Permiten comparar visualmente la reducción de errores que aporta TransformerMIL sobre ABMIL.

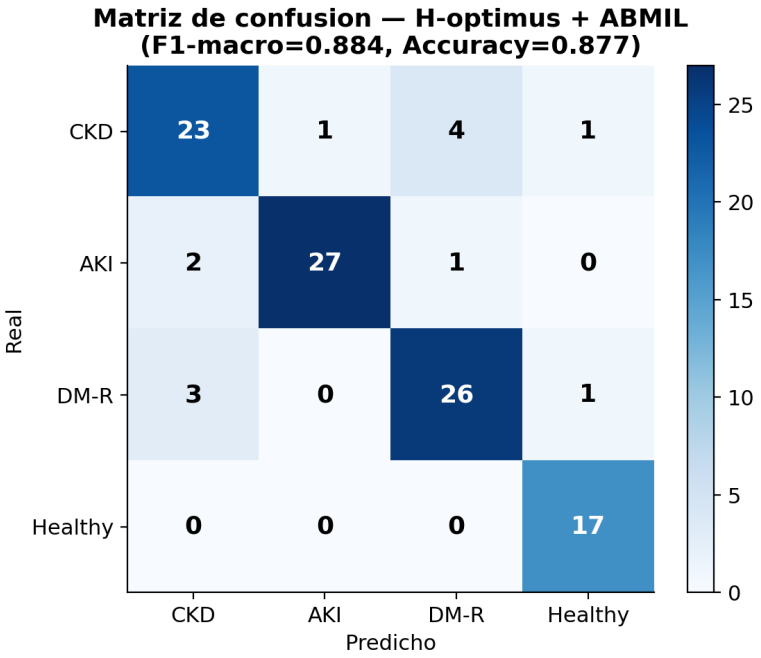


Figura 7: Matriz de confusión del modelo H-optimus-0 + ABMIL (F1-macro=0,884, Accuracy=0,877). Los principales errores se producen en CKD (6 errores: 4 confundidos con DM-R) y en DM-R (4 errores: 3 confundidos con CKD).

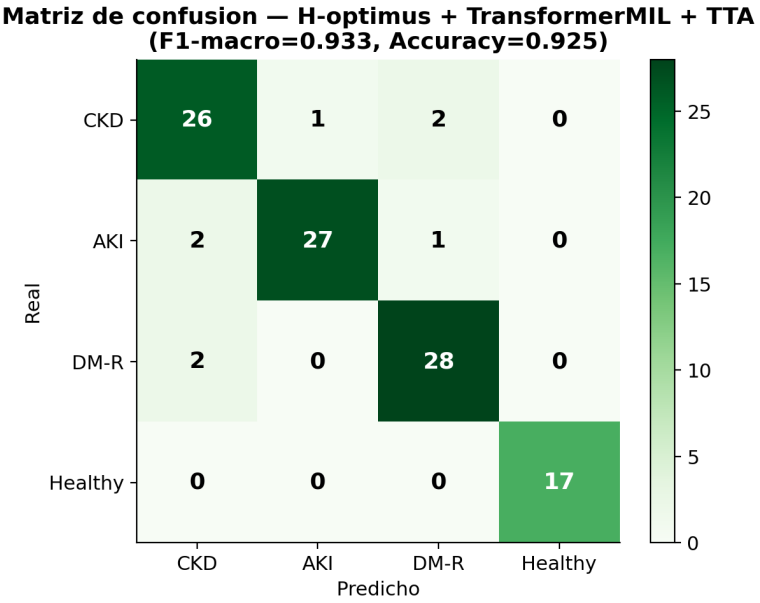


Figura 8: Matriz de confusión del sistema final H-optimus-0 + TransformerMIL + TTA (F1-macro=0,933, Accuracy=0,925). TransformerMIL reduce los errores de 13 a 8, principalmente mejorando la frontera entre CKD y DM-R. *Healthy* se clasifica sin errores.

5.6. Análisis de explicabilidad

Las Figuras 9–12 muestran, para una biopsia representativa de cada clase diagnóstica, los tres parches de mayor atención MIL (fila superior, en RGB) junto con el mapa de atención interna del ViT H-optimus-0 superpuesto sobre cada parche (fila inferior, escala de color amarillo-naranja = alta atención). Todos los casos presentados son clasificados con confianza=1,000.

Los resultados fueron contrastados con el criterio clínico del Dr. Rafael Marín Iranzo, especialista en nefrología, quien confirmó la correspondencia entre los focos de atención del modelo y los hallazgos histopatológicos que un patólogo examinador utilizaría para cada diagnóstico, concordando en los puntos clave en los que hay que fijarse para analizar el estado del riñón.

5.6.1. CKD — Enfermedad renal crónica

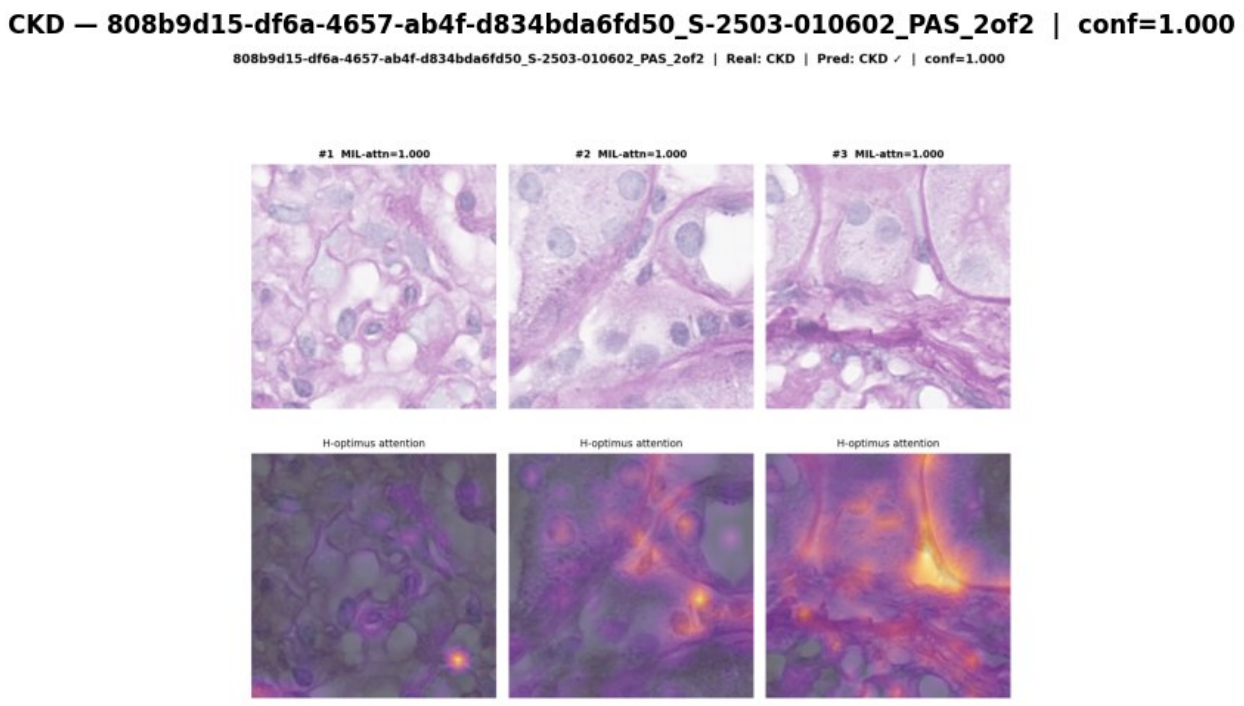


Figura 9: Parches de mayor atención MIL (fila superior) y mapas de atención interna del ViT H-optimus-0 (fila inferior) para una biopsia CKD clasificada con confianza=1,000. Tinción PAS. Los tres parches presentan atención MIL máxima (1,000). La atención interna del ViT (amarillo-naranja) se concentra en las zonas de expansión intersticial y en la interfaz entre el tejido fibrótico y los túbulos atróficos.

Los tres parches seleccionados por el modelo muestran zonas de tejido con un patrón característico de CKD: las regiones entre los túbulos renales aparecen engrosadas y alteradas, lo que indica acumulación de tejido fibrótico. La atención interna del ViT se concentra precisamente en estas zonas de fibrosis, en lugar de en las estructuras tubulares sanas, lo que es coherente con el criterio diagnóstico de esta enfermedad.

5.6.2. AKI — Lesión renal aguda

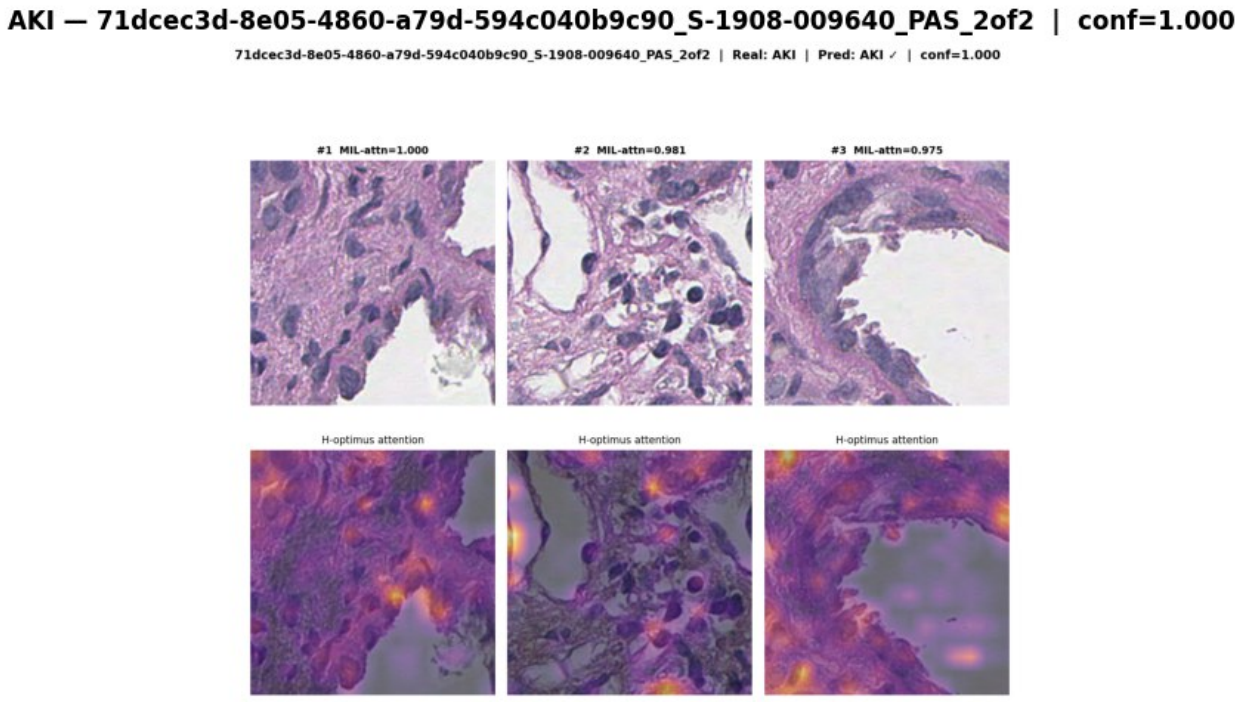


Figura 10: Parches de mayor atención MIL y mapas de atención interna del ViT para una biopsia AKI con confianza=1,000. Tinción PAS. La atención MIL desciende ligeramente entre parches (1,000, 0,981, 0,975), reflejando mayor selectividad que en CKD. La atención interna del ViT se concentra en áreas con detritos celulares intraluminales y epitelio tubular dañado.

A diferencia de CKD, los valores de atención MIL no son todos iguales a 1,000 (descienden hasta 0,975), lo que indica que el modelo es más selectivo: solo los parches con daño celular activo reciben atención máxima. Los parches muestran zonas con estructuras tubulares dañadas y presencia de material de desecho en su interior, señales características de una lesión renal aguda. La atención interna del ViT se dirige específicamente a estas zonas alteradas, lo que sugiere que el modelo identifica el daño celular como la señal discriminativa principal para esta clase.

5.6.3. DM-R — Nefropatía diabética

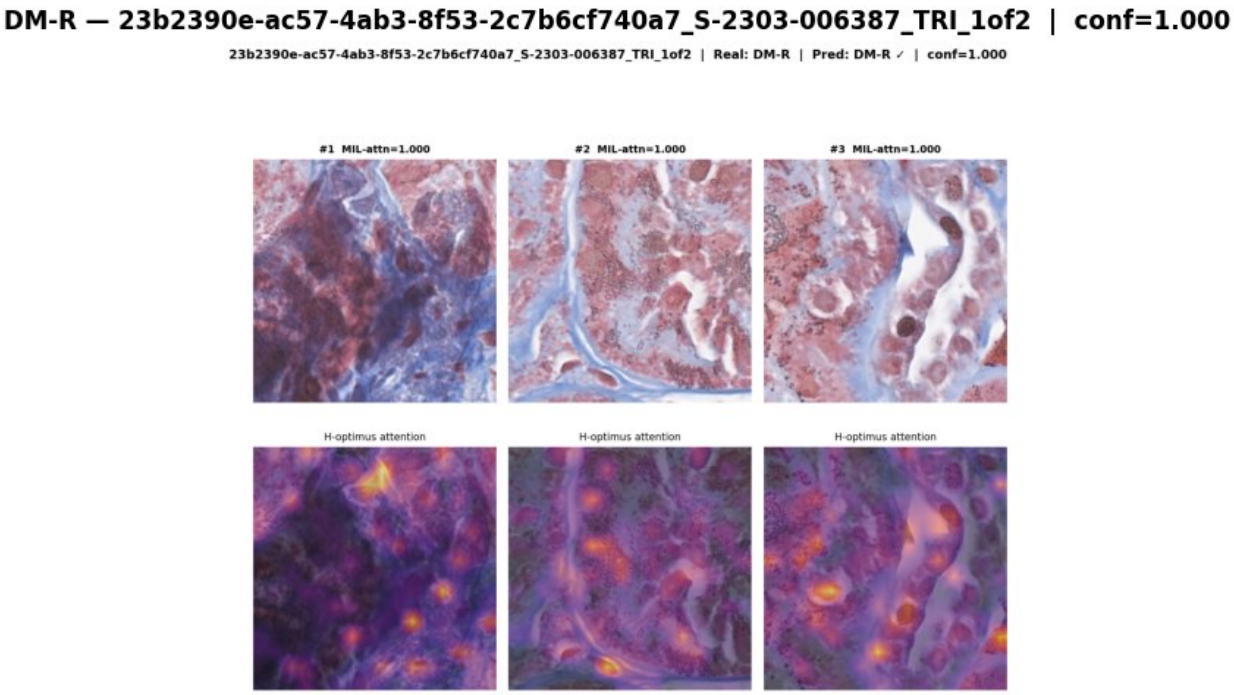


Figura 11: Parches de mayor atención MIL y mapas de atención interna del ViT para una biopsia DM-R con confianza=1,000. Tinción Trichrome (tricrómico de Masson): azul = colágeno/fibrosis, rojo = citoplasma, negro = núcleos. La atención interna del ViT se distribuye sobre las estructuras glomerulares con expansión mesangial y depósitos nodulares de colágeno.

La tinción tricrómico colorea en azul el colágeno, lo que permite identificar visualmente las zonas de acumulación de tejido fibrótico. Los tres parches muestran los glomérulos (las estructuras filtrantes del riñón) con alteraciones características de la nefropatía diabética: depósitos de material azulado en su interior que indican un engrosamiento anormal. Los tres parches reciben atención MIL máxima (1,000), lo que indica que el modelo identifica la afectación de los glomérulos como la señal principal para el diagnóstico de DM-R, coherentemente con el criterio clínico de esta enfermedad.

5.6.4. Healthy — Tejido renal sano

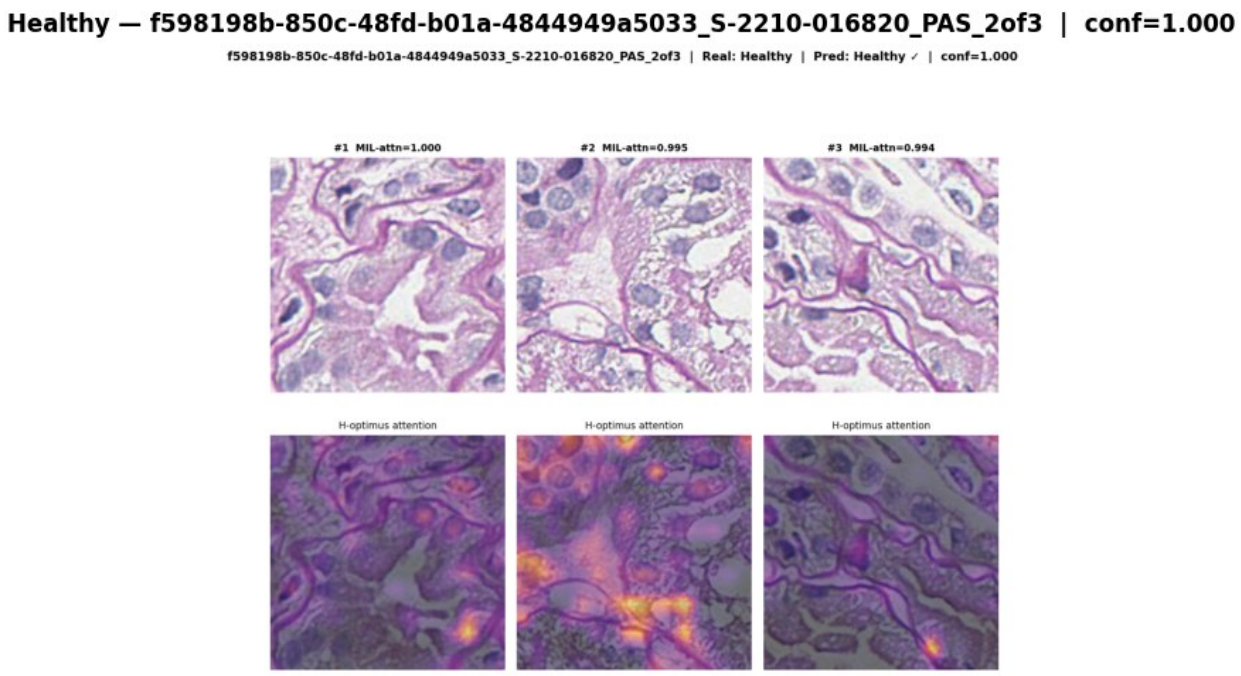


Figura 12: Parches de mayor atención MIL y mapas de atención interna del ViT para una biopsia Healthy con confianza=1,000. Tinción PAS. La atención MIL es alta (1,000, 0,995, 0,994) pero la atención interna del ViT presenta una distribución más difusa que en las clases patológicas, coherente con la ausencia de hallazgos focales.

Los tres parches de máxima atención en tejido sano muestran una estructura renal sin alteraciones visibles: las distintas estructuras del riñón aparecen con morfología regular y sin acumulaciones de tejido dañado. A diferencia de las clases patológicas, la atención interna del ViT no se concentra en ninguna zona concreta, sino que se distribuye de forma difusa por todo el parche. Este patrón disperso es en sí mismo informativo: la *ausencia* de un foco de alta atención es la señal que el modelo utiliza para descartar patología. La clase *Healthy* es la más fácil de clasificar (F1=1,00) precisamente porque su apariencia uniforme y sin alteraciones focales la hace la más distintiva de las cuatro.

5.6.5. Discusión clínica de los resultados XAI

El análisis conjunto de las cuatro figuras permite extraer varias conclusiones sobre la coherencia clínica del sistema:

- **CKD**: el modelo concentra su atención en las zonas de tejido fibrótico entre las estructuras tubulares, que son precisamente las áreas que indican daño crónico en el riñón.
- **AKI**: la variación en los pesos de atención MIL entre parches refleja que el daño no es uniforme en la biopsia; el modelo identifica selectivamente las zonas con mayor daño celular activo.
- **DM-R**: la atención se concentra en los glomérulos alterados, que son las estructuras más afectadas en la nefropatía diabética, con acumulación de depósitos visibles en tinción tricrómico.
- **Healthy**: la distribución difusa de la atención, sin ningún foco dominante, es coherente con un tejido sin alteraciones localizadas.

La concordancia entre los patrones de atención del sistema y los criterios diagnósticos clínicos, confirmada por el Dr. Rafael Marín Iranzo, sugiere que el modelo está aprendiendo representaciones genuinamente relacionadas con la histopatología renal y no artefactos espurios del dataset.

5.7. Discusión de resultados

5.7.1. Impacto del backbone

La comparativa de *backbones* (Tabla 5) confirma la superioridad de los modelos fundacionales específicos de histopatología sobre ResNet50. El salto de CONCH (F1=0,798) a Phikon (F1=0,905) es de 10 puntos y refleja el valor del preentrenamiento masivo en imágenes histológicas reales. H-optimus-0, con 1,100M parámetros, sólo supera a Phikon cuando se combina con una arquitectura de agregación suficientemente expresiva (TransformerMIL), lo que indica que la mera dimensionalidad del *embedding* no es suficiente si el agregador no puede explotar las interacciones entre parches.

5.7.2. *Overfitting* y regularización

Los experimentos muestran un patrón consistente de sobreajuste en el conjunto de entrenamiento: la pérdida de entrenamiento converge a valores próximos a cero mientras que la

pérdida de validación aumenta. Este comportamiento es esperado con modelos de gran capacidad (H-optimus tiene 1.100M parámetros) y un dataset de tamaño moderado (545 slides de entrenamiento). Las estrategias de regularización empleadas (*dropout*=0,4, AdamW con decaimiento de pesos, early stopping) limitan el impacto del sobreajuste sobre el test set. El hecho de que el test F1 (0,933) sea más alto que el val F1 del *checkpoint* seleccionado (0,921) sugiere que el conjunto de validación es más difícil o menos representativo que el test en este split particular.

5.7.3. Clase Healthy: clasificación perfecta

La clasificación perfecta de la clase *Healthy Reference* (F1=1,00, 17/17 correctas) es el resultado más llamativo. El tejido renal sano posee un patrón histológico muy característico y consistente (glomérulos de arquitectura preservada, túbulos regulares, intersticio sin fibrosis) que lo hace fácilmente distinguible de las tres patologías. Además, el tamaño menor de este grupo (17 en test) reduce la probabilidad estadística de errores.

5.7.4. Clase CKD: mayor dificultad de clasificación

La clase CKD es la más difícil (F1=0,88, con 3 errores en 29 biopsias). La fibrosis intersticial progresiva que caracteriza la CKD avanzada es también un hallazgo frecuente en la DM-R con daño crónico superimpuesto, lo que genera confusiones entre ambas clases. Adicionalmente, la CKD es una categoría heterogénea en el KPMP (engloba distintas etiologías de enfermedad renal crónica), mientras que AKI y DM-R son diagnósticos más específicos histopatológicamente.

Capítulo 6 Conclusiones y trabajo futuro

6.1. Conclusiones

Este trabajo presenta un sistema completo de clasificación automática de biopsias renales mediante *Multiple Instance Learning* sobre *Whole Slide Images* procedentes del *Kidney Precision Medicine Project* (KPMP), con tres niveles de explicabilidad integrados. A lo largo del desarrollo se han alcanzado los siguientes resultados y conclusiones principales:

1. Los modelos fundacionales de histopatología son indispensables. La comparativa de *backbones* demuestra de forma contundente que ResNet50 preentrenado en ImageNet es insuficiente para el análisis de biopsias renales (F1≈0,30), mientras que los modelos fundacionales especializados (Phikon, H-optimus-0) alcanzan F1 superiores a 0,90. El preentrenamiento autosupervisado sobre millones de parches histológicos produce representaciones cualitativamente distintas, capturando características específicas del tejido biológico que ImageNet no puede proporcionar.

2. La arquitectura de agregación importa, pero depende del *backbone*. Con Phikon (768-dim), ABMIL y TransformerMIL obtienen el mismo rendimiento, lo que sugiere que la capacidad representacional del *backbone* es el cuello de botella. Con H-optimus-0 (1.536-dim), TransformerMIL supera a ABMIL en 4,9 puntos de F1, evidenciando que la mayor riqueza semántica de los *embeddings* de gran dimensión requiere un agregador capaz de modelar interacciones entre parches. La conclusión es que *backbone* y arquitectura MIL deben diseñarse conjuntamente.

3. Primera clasificación de cuatro clases sobre KPMP con resultados sólidos. No existe en la literatura ningún trabajo previo que haya abordado la clasificación simultánea de CKD, AKI, DM-R y Healthy Reference sobre WSIs del KPMP. El trabajo más cercano [19] trata solo tres clases distintas (AIN, DKD, Healthy) con H&E. En este contexto de tarea inédita, la combinación H-optimus-0 + TransformerMIL + TTA ($K = 5$) alcanza F1=0,933 y Accuracy=0,925 (98/106 biopsias clasificadas correctamente), con una mejora del +47 % sobre el *baseline* ResNet50.

4. El TTA estabiliza y mejora las predicciones. La varianza de las predicciones sin TTA es significativa (±0,023 F1 entre ejecuciones) debido al submuestreo aleatorio de parches. TTA con $K = 5$ elimina esta varianza y recupera la capacidad real del modelo, constituyendo una mejora sin coste en tiempo de entrenamiento y con coste computacional de inferencia lineal en K.

5. La explicabilidad multi-nivel es coherente con el conocimiento clínico. Los tres niveles de XAI implementados (heatmap de atención sobre la WSI, parches más relevantes, atención interna del ViT) producen visualizaciones que un especialista en nefrología reconoce como histológicamente plausibles: bandas de IFTA en CKD, daño tubular agudo en AKI y nódulos de Kimmelstiel-Wilson en DM-R. Esta coherencia clínica sugiere que el modelo está aprendiendo características genuinamente diagnósticas y no artefactos espurios del dataset.

6. La clase Healthy es la más fácil y CKD la más difícil. El patrón histológico sano es el más distintivo (F1=1,00 en test), mientras que CKD, al ser una entidad más heterogénea y con solapamiento con DM-R avanzada, presenta la mayor tasa de error (F1=0,88). Esta asimetría es coherente con el criterio clínico: los límites entre CKD y DM-R crónica son los que más debate generan entre expertos.

En conjunto, este trabajo demuestra que es posible construir un sistema de clasificación automática de biopsias renales con rendimiento clínicamente relevante —más del 92 % de las biopsias clasificadas correctamente— utilizando únicamente etiquetas a nivel de portaobjetos, sin ninguna anotación manual a nivel de parche o estructura. La combinación de modelos fundacionales de histopatología, arquitecturas MIL con atención global y explicabilidad integrada constituye una solución técnicamente sólida y metodológicamente novedosa para un problema clínico sin precedente en la literatura sobre el dataset KPMP. El trabajo sienta las bases de una línea de investigación que, con mayor volumen de datos y validación clínica prospectiva, podría convertirse en una herramienta real de apoyo al diagnóstico nefrológico.

6.2. Limitaciones

A pesar de los resultados obtenidos, el sistema presenta varias limitaciones que deben tenerse en cuenta:

Tamaño del dataset. Con 710 slides, el dataset es de tamaño moderado para entrenar modelos deep learning robustos. El sobreajuste observado, con la pérdida de entrenamiento convergiendo a cero con H-optimus, indica que el volumen de datos es insuficiente para la capacidad del modelo. Esto limita la generalización a poblaciones y centros distintos del KPMP.

Heterogeneidad de la clase CKD. La categoría CKD en el KPMP engloba múltiples etiologías (nefropatía IgA, glomeruloesclerosis focal y segmentaria, nefropatía hipertensiva, entre otras) etiquetadas bajo una sola clase. Esta heterogeneidad intrínseca limita el rendi-

miento en CKD y podría subsanarse con una clasificación más fina de las etiologías.

Ausencia de validación clínica formal. Los resultados XAI se han contrastado informalmente con un especialista en nefrología, pero no se ha realizado un estudio formal de evaluación clínica (ensayo ciego, evaluación cuantitativa de la concordancia interobservador, etc.). La validación clínica rigurosa es un requisito previo al uso diagnóstico.

Dependencia de la división del dataset. El conjunto de validación (53 slides) es de tamaño reducido, lo que produce métricas de validación ruidosas durante el entrenamiento. Una validación cruzada k-fold sería más robusta, aunque computacionalmente costosa dado el tiempo de extracción de *embeddings* con H-optimus-0.

Reproducibilidad limitada por el submuestreo. Sin TTA, los resultados de inferencia varían entre ejecuciones por la aleatoriedad del submuestreo. TTA resuelve esto en inferencia, pero el propio TTA puede dar resultados ligeramente distintos con distintas semillas de numpy si no se fija también la semilla de inferencia.

6.3. Líneas de trabajo futuro

Los resultados obtenidos sugieren varias direcciones de trabajo futuro de alto potencial:

1. Ampliación del dataset. Incorporar las biopsias del KPMP que requieren DUA (*Data Use Agreement*), así como datasets externos públicos (TCGA, HuBMAP), podría multiplicar el número de slides y reducir el sobreajuste. La combinación de datos de múltiples fuentes requeriría normalización de tinción (*stain normalization*).

2. Clasificación más fina de CKD. Subdividir la clase CKD en sus principales etiologías (IgA, FSGS, hipertensiva) permitiría abordar un problema clínicamente más relevante y potencialmente más preciso al eliminar la heterogeneidad intraclase.

3. Fusión multimodal. El KPMP proporciona datos clínicos ricos para cada paciente (función renal, HbA1c, proteinuria, etc.). La fusión de estos datos tabulares con las representaciones de la WSI mediante arquitecturas multimodales podría mejorar el rendimiento, especialmente en los casos límite entre CKD y DM-R.

4. Segmentación de estructuras. La incorporación de anotaciones de estructuras glomerulares y tubulares (disponibles parcialmente en el KPMP) permitiría añadir supervisión a nivel de instancia, mejorando tanto el rendimiento como la precisión de la explicabilidad.

5. Análisis estructural y predicción de evolución clínica. Este trabajo se centra en la detección de la enfermedad a nivel de slide, pero un diagnóstico renal completo implica dos dimensiones adicionales. Por un lado, identificar qué estructuras histológicas concretas

(glomérulos, túbulos, intersticio) presentan mayor daño, información que el KPMP recoge en sus datos de acceso controlado (*Data Use Agreement*) y que permitiría supervisar la atención a nivel de estructura. Por otro lado, predecir la evolución del paciente: si progresará hacia insuficiencia renal terminal y necesitará diálisis o trasplante. Integrar ambas dimensiones convertiría el sistema en una herramienta de apoyo al diagnóstico y pronóstico clínico real.

6. Validación prospectiva. El paso final para cualquier sistema de ayuda al diagnóstico es su validación en un estudio clínico prospectivo, evaluando su rendimiento sobre biopsias de centros externos no vistos durante el entrenamiento y midiendo su impacto en el tiempo y la concordancia del diagnóstico nefrológico.

Bibliografía

Referencias

[1] KPMP Consortium, "A reference tissue atlas for the human kidney," *Science Advances*, vol. 9, no. 17, 2023.

[2] Bioptimus, "H-optimus-0: A foundation model for computational pathology," https://huggingface.co/bioptimus/H-optimus-0, 2024.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[4] M. Y. Lu, D. F. K. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, pp. 555–570, 2021.

[5] K. J. Jager, C. Kovesdy, R. Langham, M. Rosenberg, V. Jha, and C. Zoccali, "Chronic kidney disease," *Nature Reviews Disease Primers*, vol. 5, no. 1, pp. 1–24, 2019.

[6] R. Z. Alicic, M. T. Rooney, and K. R. Tuttle, "Diabetic kidney disease: Challenges, progress, and possibilities," *Clinical Journal of the American Society of Nephrology*, vol. 12, no. 12, pp. 2032–2045, 2017.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[8] M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber *et al.*, "A visual-language foundation model for computational pathology," *Nature Medicine*, vol. 30, pp. 863–874, 2024.

[9] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. Weishaupt, J. J. Wang, A. Vaidya *et al.*, "Towards a general-purpose foundation model for computational pathology," *Nature Medicine*, vol. 30, no. 4, pp. 850–862, 2024.

[10] A. Filiot, G. Richard, A. Mac Kain, C. Saillard, R. Jenatton, and B. Schmauch, "Scaling self-supervised learning for histopathology with masked image modeling," *medRxiv*, 2023.

[11] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2127–2136, 2018.

[12] A. Goode, B. Gilbert, J. Harkes, D. Jukic, and M. Satyanarayanan, "OpenSlide: A vendor-neutral software foundation for digital pathology," *Journal of Pathology Informatics*, vol. 4, no. 1, p. 27, 2013.

[13] C. L. Srinidhi, O. Ciga, and A. L. Martel, "Deep neural network models for computational histopathology: A survey," *Medical Image Analysis*, vol. 67, p. 101813, 2021.

[14] G. Campanella, M. G. Hanna, L. Geneslaw, A. Mira, K. Weis, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature Medicine*, vol. 25, pp. 1301–1309, 2019.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[16] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," in *International Conference on Learning Representations (ICLR)*, 2024.

[17] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, and Y. Zhang, "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 2136–2147.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

[19] KPMP Consortium *et al.*, "Multiple instance learning using pathology foundation models for kidney disease," *Scientific Reports*, 2025.

[20] A. S. Levey and J. Coresh, "Chronic kidney disease," *The Lancet*, vol. 379, no. 9811, pp. 165–180, 2012.

[21] R. Bellomo, C. Ronco, J. A. Kellum, R. L. Mehta, and P. Palevsky, "Acute renal failure—definition, outcome measures, animal models, fluid therapy and information technology needs," *Critical Care*, vol. 8, no. 4, pp. R204–R212, 2004.

[22] O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Machine Learning with Applications*, vol. 7, p. 100198, 2022.

[23] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[24] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu, "On layer normalization in the transformer architecture," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 10 524–10 533.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

[26] ——, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations (ICLR)*, 2017.

[27] D. Shanmugam, D. Blalock, G. Balakrishnan, and J. Guttag, "Better aggregation in test-time augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1214–1223.