Pontificia Universidad Javeriana Cali
Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos
Proyecto Aplicado III

Diseño e implementación de modelos de Aprendizaje de Instancias Múltiples para la clasificación débilmente supervisada de imágenes histopatológicas de cáncer de próstata

Juan José Restrepo Rosero
María Valentina Belalcázar Perdomo

Director: Dr. Julian Gil González

24 de febrero de 2026

Pontificia Universidad
JAVERIANA
Cali

Nota de Aceptación

Aprobado por el Comité de Trabajo de Grado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana para optar al título de Magíster en Ciencia de Datos.

Dr. Diego Luis Linares Ospina
Director Maestría en Ciencia de Datos

Dr. Julian Gil González
Director Trabajo de Grado

Dr. Nombre Apellido
Jurado

Dr. Nombre Apellido
Jurado

Santiago de Cali, 24 de febrero de 2026

Señores
Pontificia Universidad Javeriana – Cali

Facultad de Ingeniería y Ciencias
Cali

Cordial Saludo.

Nos permitimos presentar a su consideración el Proyecto Aplicado titulado "Diseño e implementación de modelos de Aprendizaje de Instancias Múltiples para la clasificación débilmente supervisada de imágenes histopatológicas de cáncer de próstata" con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el proyecto de grado y posteriormente optar al título de Magíster en Ciencia de Datos.

Al firmar aquí, damos fe que entendemos y conocemoslas directrices para la presentación de trabajos de grado de la Facultad de Ingeniería y Ciencias aprobadas el 26 de noviembre de 2009, donde se establecen los plazos y normas para el desarrollo del anteproyecto y del trabajo de grado.
Atentamente,

Juan José Restrepo Rosero
Código: 8939280

María Valentina Belalcázar Perdomo
Código: 8992284

Santiago de Cali, 24 de febrero de 2026

Señores
Pontificia Universidad Javeriana – Cali

Facultad de Ingeniería y Ciencias
Cali

Cordial Saludo.

Por medio de la presente me permito informarle que los estudiantes de la Maestría en Ciencia de Datos Juan José Restrepo Rosero (cod: 8939280) y María Valentina Belalcázar Perdomo (cod: 8992284) trabajaron bajo mi dirección en el Proyecto Aplicado titulado 'Diseño e implementación de modelos de Aprendizaje de Instancias Múltiples para la clasificación débilmente supervisada de imágenes histopatológicas de cáncer de próstata'.

Atentamente,

_____

Dr. Julian Gil González
Director Trabajo de Grado

Dedicatoria

El presente trabajo de grado se lo dedicamos a Dios, familiares y a amigos cercanos que nos han acompañado en esta experiencia de formación como magisters en ciencia de datos.

Agradecimientos

Expresamos nuestro agradecimiento al Dr. Julian Gil González, director de este proyecto, cuya entrega y compromiso incondicional fueron los cimientos sobre los cuales se construyó esta investigación. Más allá de su orientación experta y técnica, su seguimiento constante y su disposición para el diálogo constructivo fueron fundamentales para navegar la complejidad de este trabajo. Su guía no solo nos permitió alcanzar los objetivos académicos planteados, sino que se convirtió en una fuente de inspiración y resiliencia para superar los desafíos inherentes al aprendizaje profundo y la ciencia de datos, fomentando en nosotros un pensamiento crítico y riguroso.

Asimismo, extendemos nuestra gratitud al cuerpo docente de la Maestría en Ciencia de Datos de la Pontificia Universidad Javeriana Cali. Cada cátedra, debate y lección compartida en las aulas ha contribuido a nuestra formación integral. Su rigor académico y la pasión por transmitir el conocimiento han sido los pilares que nos permitieron consolidar las competencias técnicas y analíticas necesarias para alcanzar este título profesional. Gracias por exigirnos excelencia y por proporcionarnos las herramientas para enfrentar los retos del mundo tecnológico actual.

Finalmente, dedicamos este logro con amor y gratitud infinita a nuestras familias. Ustedes han sido nuestro refugio y fortaleza a lo largo de este exigente camino; gracias por su apoyo inagotable, por la sabiduría de sus consejos y por ser ese soporte emocional incondicional que nos impulsó a perseverar en los momentos de incertidumbre. Su presencia, sacrificio y confianza ciega en nuestras capacidades han sido el motor fundamental y la razón de ser de este esfuerzo. La culminación exitosa de este trabajo de grado es, en esencia, un triunfo compartido con ustedes.

Resumen

Este trabajo de grado presenta el diseño, implementación y evaluación de un pipeline experimental reproducible basado en Aprendizaje de Instancias Múltiples (Multiple Instance Learning, MIL) para la clasificación débilmente supervisada de imágenes de lámina completa (Whole Slide Images, WSI) en cáncer de próstata. El estudio aborda dos desafíos fundamentales de la histopatología digital: la variabilidad interobservador en la gradación tumoral y la escasez de anotaciones locales a nivel de parche.

La metodología propuesta integra un esquema de preprocesamiento orientado a la selección de regiones tisulares relevantes mediante filtrado en el espacio de color HSV, extracción de representaciones profundas utilizando una ResNet-50 preentrenada como extractor congelado de características, y entrenamiento de arquitecturas MIL con mecanismos de agregación clásicos y basados en atención. El proceso experimental se desarrolló sobre la base de datos pública SICAPv2 y se evaluó mediante validación cruzada estricta estratificada por paciente (*GroupKFold*), garantizando la separación completa entre entrenamiento y prueba a nivel de WSI.

Los resultados evidencian que los modelos basados en atención, particularmente SmABMIL, superan consistentemente a los enfoques de pooling tradicional en métricas clínicas relevantes a nivel de lámina completa. En promedio, se obtuvo un F1-score superior a 0.83 y una AUC-ROC cercana a 0.86, junto con una mayor estabilidad inter-fold y mejor equilibrio entre sensibilidad y especificidad. Adicionalmente, la incorporación de mecanismos de atención permitió generar mapas de relevancia espacial coherentes con patrones histopatológicos asociados a malignidad, fortaleciendo la interpretabilidad del sistema.

En conjunto, el trabajo demuestra la viabilidad del aprendizaje con supervisión débil para clasificación de cáncer prostático en WSI y establece una base metodológica sólida y reproducible para el desarrollo de sistemas de apoyo a la decisión clínica en patología digital.

Palabras clave: Aprendizaje de instancias múltiples (MIL); Whole Slide Images (WSI); Histopatología digital; Cáncer de próstata; Redes neuronales convolucionales; Interpretabilidad.

Abstract

This undergraduate thesis presents the design, implementation, and evaluation of a reproducible experimental pipeline based on Multiple Instance Learning (MIL) for weakly supervised classification of Whole Slide Images (WSI) in prostate cancer. The study addresses two central challenges in digital histopathology: interobserver variability in tumor grading and the limited availability of detailed patch-level annotations.

The proposed methodology integrates a preprocessing stage focused on relevant tissue region selection using HSV color space filtering, deep feature extraction through a frozen ResNet-50 backbone, and the training of MIL architectures employing both classical pooling and attention-based aggregation mechanisms. The experimental framework was developed using the public SICAPv2 dataset and evaluated through strict patient-level cross-validation (*GroupKFold*), ensuring complete separation between training and testing WSIs.

The results demonstrate that attention-based MIL models, particularly SmABMIL, consistently outperform traditional pooling strategies in clinically relevant slide-level metrics. On average, the models achieved an F1-score above 0.83 and an AUC-ROC close to 0.86, along with improved inter-fold stability and a better balance between sensitivity and specificity. Furthermore, the incorporation of attention mechanisms enabled the generation of spatial relevance maps aligned with histopathological patterns associated with malignancy, enhancing the interpretability of the system.

Overall, this work validates the feasibility of weakly supervised learning for prostate cancer classification in WSI and establishes a solid, reproducible methodological foundation for the development of clinically oriented decision-support systems in digital pathology.

Keywords: Multi-instance learning (MIL); Whole Slide Images (WSI); Digital histopathology; Prostate cancer; Convolutional neural networks; Interpretability.

Índice general

Índice de figuras

Capítulo 1

Introducción

Las técnicas de diagnóstico asistido por inteligencia artificial en histopatología digital han demostrado un notable potencial para mejorar la precisión, reproducibilidad y eficiencia del análisis de tejido, contribuyendo a la reducción de la variabilidad interobservador y de los tiempos de informe clínico [?]. En el caso del cáncer de próstata, una de las neoplasias más prevalentes a nivel mundial, el análisis histopatológico se mantiene como el procedimiento de referencia estándar para determinar la presencia y agresividad tumoral mediante la escala de Gleason. Sin embargo, este proceso depende en gran medida de la experiencia del patólogo, lo que puede introducir subjetividad y variabilidad diagnóstica [?].

La digitalización de láminas histológicas completas en imágenes de alta resolución, conocidas como *Whole Slide Images* (WSI), ha abierto nuevas oportunidades para el uso de modelos de aprendizaje profundo en patología computacional. No obstante, estas imágenes suelen estar etiquetadas únicamente a nivel de muestra (*slide-level*) y no a nivel de parche o región (instancia), lo que dificulta la aplicación directa de enfoques completamente supervisados que requieren anotaciones finas [?]. Frente a esta limitación, el paradigma de aprendizaje de instancias múltiples (*Multiple Instance Learning*, MIL) surge como una alternativa práctica y eficiente, debido a que permite entrenar modelos utilizando etiquetas globales por WSI, reduciendo significativamente la carga de anotación experta sin renunciar a un rendimiento clínicamente relevante [?][?].

En este contexto, el presente proyecto propone el diseño de un pipeline modular y reproducible para la aplicación de técnicas MIL sobre imágenes histopatológicas de cáncer de próstata, utilizando el dataset público SICAPv2 [?]. El pipeline integra etapas de preprocesamiento y extracción de parches, extracción de representaciones mediante redes neuronales convolucionales preentrenadas y módulos de agregación débilmente supervisados basados en mecanismos de atención. Se pone especial énfasis en la reproducibilidad experimental, el versionado de datos y la validación cruzada estratificada por paciente, siguiendo recomendaciones ampliamente aceptadas en patología computacional [?] [?].

La evaluación del sistema se realiza mediante métricas estándar como AUC, precisión, *recall* y F1-score, complementadas con análisis estadísticos que permiten comparar distintas variantes del modelo más allá de cifras puntuales. Adicionalmente, se estudia en profundidad el comportamiento del mecanismo de atención, con el objetivo de aportar interpretabilidad y analizar si las instancias con mayor peso de atención corresponden a regiones histopatológicamente relevantes, un aspecto crítico para la confianza y la posible adopción clínica de este tipo de modelos.

Como resultado de esta investigación, se espera obtener un conjunto de parches reproducible y versionado derivado del dataset SICAPv2, implementaciones de código abierto de distintas variantes de MIL y extractores de características, así como un estudio comparativo riguroso que analice los compromisos entre desempeño, generalización e interpretabilidad. En conjunto, estos aportes buscan proporcionar evidencia técnica y clínica sólida que respalde el uso de herramientas de apoyo al diagnóstico basadas en aprendizaje profundo en el ámbito de la histopatología digital.

Capítulo 2

Descripción del Problema

El cáncer de próstata constituye una de las principales causas de morbilidad y mortalidad en la población masculina a nivel mundial. Según estimaciones recientes de la Organización Mundial de la Salud (OMS) y del Global Cancer Observatory (GCO), se reportan anualmente más de 1.4 millones de casos nuevos, con una carga especialmente elevada en países de ingresos medios y bajos debido a barreras en el acceso al diagnóstico temprano y al tratamiento oportuno [?]. En el contexto colombiano, esta enfermedad figura entre las primeras causas de muerte por cáncer en hombres, lo que subraya la necesidad de fortalecer estrategias de detección precoz y estratificación de riesgo que apoyen la toma de decisiones clínicas [?].

El diagnóstico histopatológico mediante la escala de Gleason constituye el pilar de la estratificación pronóstica; sin embargo, su aplicación está sujeta a una variabilidad interobservador significativa. Esta subjetividad es especialmente crítica en la discriminación entre patrones limítrofes (por ejemplo, Gleason 3 frente a Gleason 4) y en la identificación de subtipos histológicos asociados a peor pronóstico, como el patrón *cribriforme*. Dicho patrón se caracteriza por formaciones glandulares con arquitectura tipo criba y se asocia a mayor agresividad tumoral y riesgo de recurrencia, por lo que su correcta detección tiene implicaciones terapéuticas relevantes [?] [?].

La digitalización de muestras histopatológicas y la generación de *Whole Slide Images* (WSI) han habilitado el análisis computacional del tejido completo, pero introducen retos técnicos sustanciales. Entre ellos se incluyen el tamaño gigapíxel de las imágenes, la heterogeneidad morfológica intratumoral, las variaciones de tinción entre centros y la presencia de artefactos derivados de la preparación y el escaneo de las muestras [?]. Adicionalmente, la disponibilidad de conjuntos de datos con anotaciones detalladas a nivel de parche (*patch-level*) es limitada, debido al alto costo y tiempo requeridos para su generación por expertos.

Es importante aclarar la diferencia entre los niveles de anotación que condicionan la metodología de aprendizaje. Una etiqueta a nivel de lámina/muestra (*slide-level*) asigna una única etiqueta/clase a toda la WSI, mientras que una etiqueta a nivel de instancia (*patch-level*) identifica la clase de cada parche individual. La ausencia generalizada de anotaciones *patch-level* ha motivado el uso de enfoques de aprendizaje débil, particularmente el paradigma de *Multiple Instance Learning*, que permite entrenar modelos utilizando únicamente etiquetas globales a nivel de WSI.

Desde la perspectiva de la ciencia de datos, la combinación de imágenes de gran escala, supervisión débil y heterogeneidad interinstitucional genera un conjunto de desafíos metodológicos bien definidos. Estos incluyen la necesidad de estrategias eficientes de preprocesamiento y representación, el manejo del desbalance de clases, la evaluación robusta del desempeño y la mitigación de *domain shifts* derivados de diferencias en protocolos de tinción y adquisición [?][?]. Asimismo, la escasez de subtipos relevantes desde el punto de vista clínico plantea retos adicionales para el aprendizaje y la generalización de los modelos.

En este contexto, el problema central que aborda este proyecto consiste en diseñar y evaluar un pipeline de aprendizaje débil capaz de clasificar imágenes histopatológicas prostáticas a partir de etiquetas *slide-level*,

preservando información diagnóstica relevante y garantizando una evaluación experimental reproducible. Particularmente, se investiga el uso de variantes de MIL basadas en mecanismos de atención, con el fin de analizar no solo su desempeño cuantitativo, sino también su estabilidad y potencial interpretativo.

Finalmente, el proyecto adopta prácticas esenciales de reproducibilidad, tales como particiones estratificadas por paciente, control de semillas aleatorias y versionado del conjunto de parches, que permiten realizar comparaciones internas consistentes entre distintas configuraciones del modelo. Evaluaciones más extensas, como validaciones interinstitucionales o análisis cualitativos con múltiples patólogos, se reconocen como relevantes, pero se consideran fuera del alcance de este proyecto y se proponen como líneas de trabajo futuro.

Capítulo 3

Objetivos

3.1. Objetivo General

Desarrollar un modelo de visión por computador a partir de técnicas de Aprendizaje con Múltiples Instancias para la clasificación débilmente supervisada de imágenes de histopatología de cáncer de próstata, con el fin de maximizar la precisión diagnóstica y facilitar su interpretación clínica.

3.2. Objetivos Específicos

1. Implementar técnicas de procesamiento de imágenes de histopatología prostática, incluyendo corrección de color, eliminación de artefactos y normalización de tinciones.
2. Implementar modelos de visión por computador basados en Aprendizaje de Instancias Múltiples con Deep Learning para la clasificación de Whole Slide Images.
3. Evaluar el rendimiento de los modelos desarrollados mediante métricas estándar (AUC, precisión, recall, F1-score y coeficiente Kappa) con validación cruzada.

Capítulo 4

Marco Teórico

4.1. Términos clave

A continuación se presentan las definiciones de los términos centrales utilizados a lo largo de este trabajo, con el fin de establecer una base conceptual precisa y homogénea que facilite la comprensión de los capítulos posteriores.

1. **Bag e instancia:** En el paradigma de *Aprendizaje de Instancias Múltiples* (Multiple Instance Learning, MIL), un *bag* corresponde a un conjunto de instancias individuales que no cuentan con etiquetas propias, pero que están asociadas colectivamente a una etiqueta de nivel superior [**?**]. Formalmente, el aprendizaje se realiza a partir de pares $(\mathcal{X}_i, y_i)$, donde $\mathcal{X}_i$ denota el conjunto de instancias y y_i la etiqueta del bag. En el contexto de la histopatología digital, cada *Whole Slide Image* (WSI) se modela como un bag, mientras que los parches extraídos de dicha imagen constituyen las instancias. Esta formulación resulta especialmente adecuada en escenarios clínicos reales, donde las anotaciones detalladas a nivel de parche suelen ser inexistentes o costosas de obtener.

2. **Técnicas de agregación:** Corresponden a los mecanismos mediante los cuales la información proveniente de múltiples instancias dentro de un bag es combinada para producir una predicción global. Estas técnicas definen explícitamente la función de decisión del modelo MIL. Entre los enfoques más comunes se encuentran los esquemas clásicos de *pooling*, como *mean pooling* y *max pooling*, que actúan como líneas base, así como métodos más avanzados basados en mecanismos de atención, los cuales aprenden a ponderar dinámicamente la contribución de cada instancia en función de su relevancia para la tarea de clasificación [**?**][**?**] [**?**].

3. **Supervisión débil:** Hace referencia a escenarios de aprendizaje en los que la información de etiquetado es incompleta, imprecisa o disponible únicamente a un nivel de granularidad superior, en contraste con la supervisión totalmente anotada a nivel de instancia [**?**] [**?**] [**?**]. En el caso de la clasificación de WSIs, las etiquetas suelen estar disponibles únicamente a nivel de lámina completa, lo que impone restricciones significativas al entrenamiento supervisado tradicional y motiva el uso de enfoques MIL como una solución viable y clínicamente realista.

4. **Mecanismo de atención:** Es una estrategia de agregación que permite al modelo aprender pesos diferenciados para cada instancia dentro de un bag, reflejando su contribución relativa a la predicción final. En el contexto de MIL, los mecanismos de atención no solo mejoran la robustez del modelo frente a ruido y heterogeneidad intra-bag, sino que además proporcionan interpretabilidad intrínseca al evidenciar qué subconjuntos de instancias resultan más informativos para la decisión del modelo, lo cual es particularmente relevante en aplicaciones médicas [**?**] [**?**].

5. **Heterogeneidad intra-lámina:** Describe la variabilidad estructural, morfológica y semántica presente entre los parches que componen una misma WSI. Esta heterogeneidad puede deberse a la coexistencia de tejido sano, regiones tumorales de distinta agresividad, artefactos de preparación o zonas no

informativas, y constituye uno de los principales desafíos en la clasificación automática de imágenes histopatológicas [?].

6. **Embeddings o representaciones profundas:** Son vectores de características de dimensión reducida que codifican información semántica relevante extraída por redes neuronales convolucionales (CNNs) entrenadas como extractores de características. En este trabajo, cada parche es representado mediante un embedding en un espacio latente, lo que permite desacoplar la etapa de extracción de características de la etapa de agregación y clasificación basada en MIL [?].

7. **ISUP y escala de Gleason:** La *International Society of Urological Pathology* (ISUP) define un sistema de gradación del cáncer de próstata basado en la evaluación de patrones arquitecturales glandulares observados en tinciones hematoxilina-eosina (H&E) [?] [?]. La escala de Gleason asigna puntuaciones que reflejan la agresividad tumoral, mientras que las categorías ISUP agrupan dichas puntuaciones en clases clínicamente significativas. En este trabajo, estas categorías son reformuladas en un esquema binario con el objetivo de abordar una tarea de clasificación alineada con decisiones diagnósticas prácticas y con la disponibilidad de datos.

8. **Coeficiente de concordancia de Cohen (Kappa):** Es una métrica estadística utilizada para cuantificar el grado de acuerdo entre dos clasificadores discretos, corrigiendo por la probabilidad de coincidencia atribuible al azar. Esta medida se define como

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

donde p_o es la proporción de acuerdos observados y p_e es la proporción de acuerdos esperados por azar. El coeficiente Kappa toma valores entre -1 y 1, donde valores cercanos a 1 indican un alto nivel de acuerdo, valores cercanos a 0 sugieren acuerdo consistente con el azar, y valores negativos representan acuerdo menor al esperado por azar [?]. Esta métrica se emplea frecuentemente en estudios clínicos y de clasificación para evaluar la concordancia entre predicciones automatizadas y anotaciones de referencia.

4.2. Cáncer de próstata y escala de Gleason

El cáncer de próstata es una de las neoplasias más frecuentes y una de las principales causas de morbilidad y mortalidad en la población masculina a nivel mundial [?] [?]. La detección temprana y precisa de esta enfermedad es fundamental para orientar decisiones terapéuticas y estrategias de manejo individualizado, debido a la amplia variabilidad biológica de la enfermedad y su impacto en la supervivencia a largo plazo.

El diagnóstico clínico convencional se basa en el análisis histopatológico de biopsias teñidas con hematoxilina y eosina (H&E), donde el patólogo examina patrones arquitecturales del tejido prostático bajo el microscopio. Este examen produce la asignación de un **puntaje de Gleason**, un sistema de gradación que clasifica la agresividad tumoral con base en la arquitectura glandular observada en las muestras [?] [?].

La escala de Gleason combina dos patrones arquitecturales, primario y secundario, para obtener un puntaje sumado que refleja la agresividad biológica del tumor.

- Puntajes bajos (por ejemplo, Gleason 6) se asocian con tumores bien diferenciados y menor agresividad.
- Puntajes altos (Gleason 8–10) indican tumores más indiferenciados y de mayor agresividad clínica.

Esta información se resume y agrupa en categorías clínicamente relevantes definidas por la **International Society of Urological Pathology (ISUP)** [?] [?].

Ahora bien, un patrón histológico particularmente desafiante es el **cribriforme**, caracterizado por estructuras glandulares complejas que se asocian con peor pronóstico y mayor riesgo de recurrencia. Numerosos estudios han documentado que la presencia de patrones cribiformes se correlaciona con desenlaces clínicos adversos, lo que subraya la importancia de su identificación precisa [?]. La Figura 4.1 muestra ejemplos de este patrón histológico, evidenciando la complejidad de su arquitectura glandular y la dificultad diagnóstica que representa.

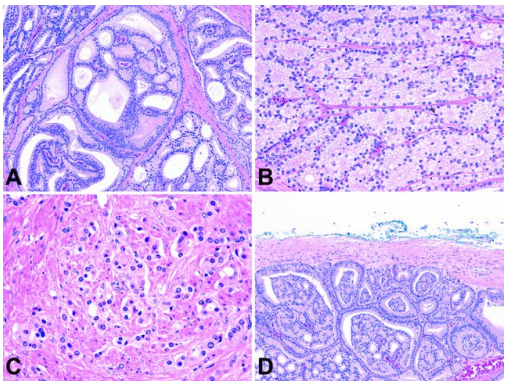

Figura 4.1: Ejemplos representativos de patrones cribiformes (Gleason 4) en tejido prostático [?].

A pesar de la importancia clínica de la gradación de Gleason, su asignación presenta variabilidad **interobservador** significativa, incluso entre patólogos experimentados [?] [?] [?]. Dicha variabilidad es especialmente notoria en la distinción entre **Gleason 3 y Gleason 4** y en la identificación de patrones con riesgo clínico intermedio o alto, lo que genera retos en la reproducibilidad y consistencia diagnóstica.

Para integrar esta complejidad diagnóstica con métodos computacionales basados en aprendizaje automático, es frecuente reformular el problema en términos clínicos significativos. Una aproximación clínicamente respaldada consiste en **binarizar las categorías ISUP** en dos grupos de interés:

- **Bajo grado**: tumores con menor agresividad (ISUP grados 1–2).
- **Alto grado**: tumores con mayor probabilidad de progresión clínica adversa (ISUP grados ≥ 3).

Esta reformulación reduce la ambigüedad diagnóstica en casos limítrofes y se alinea con decisiones clínicas reales, tales como la elección entre tratamientos activos frente a intervenciones más agresivas. Diversos trabajos en patología computacional han adoptado esquemas similares cuando el objetivo es la **clasificación binaria de riesgo clínico** en cáncer de próstata [?] [?].

La digitalización de imágenes histopatológicas (WSIs) genera grandes volúmenes de información visual de alta resolución [?]. Procesar y analizar estas imágenes exhaustivamente supera las capacidades humanas en términos de rapidez y consistencia, especialmente frente a:

- Variabilidad morfológica intrínseca del tejido,
- Artefactos de preparación y tinción,

- La elevada cantidad de parches contenidos en una sola WSI.

Estas limitaciones han motivado la adopción de técnicas computacionales avanzadas, particularmente en aprendizaje profundo, para apoyar la evaluación histopatológica y mejorar la reproducibilidad diagnóstica. Las arquitecturas de **Multiple Instance Learning** (MIL), junto con mecanismos de atención, se presentan como soluciones naturales para integrar supervisión débil con representación de patrones discriminativos en WSIs.

En consecuencia, el desarrollo de modelos computacionales capaces de imitar y fortalecer el juicio clínico en la gradación de Gleason constituye un área de investigación activa y de alto impacto clínico. La variabilidad interobservador, la complejidad morfológica de ciertos patrones histológicos y la necesidad de evaluaciones reproducibles han impulsado el uso de enfoques basados en aprendizaje profundo para apoyar el diagnóstico histopatológico.

4.3. Histopatología digital y Whole Slide Images (WSI)

La histopatología digital constituye una evolución significativa de la patología convencional mediante la captura digital de preparaciones histológicas completas en forma de *Whole Slide Images* (WSIs). Estas representaciones digitales de resolución gigapíxel preservan de forma íntegra la arquitectura tisular y habilitan su análisis sistemático mediante algoritmos computacionales avanzados, apoyando criterios diagnósticos más objetivos y consistentes que el examen microscópico manual tradicional [?] [?].

Las WSIs se generan mediante el escaneo de portaobjetos histológicos teñidos (usualmente con hematoxilina y eosina), produciendo imágenes digitales completas de los cortes de tejido que pueden almacenarse, visualizarse y compartirse con alta fidelidad. Este flujo de trabajo permite la integración de herramientas informáticas que asisten en la identificación de regiones relevantes, la cuantificación de características morfológicas y la generación de métricas automatizadas para apoyar la decisión diagnóstica clínica [?].

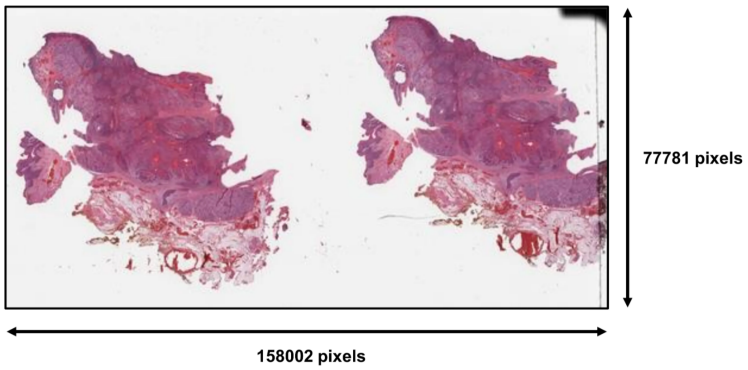

Figura 4.2: Ejemplo representativo de una imagen de alta resolución WSI que mantiene la arquitectura tisular completa [?].

Las WSIs pueden alcanzar resoluciones de varios gigapíxeles, lo que las hace extremadamente grandes para el procesamiento directo mediante modelos de visión por computador tradicionales. Esta característica introduce importantes desafíos técnicos:

- **Tamaño de imagen extremo:** El tamaño total de una WSI excede con creces la capacidad de entrada de redes neuronales convencionales, lo que impide el análisis de toda la imagen de una sola vez.
- **Heterogeneidad visual:** Las áreas dentro de una misma WSI pueden contener tejido sano, regiones tumorales de distintos grados, artefactos de preparación y variación de tinción entre laboratorios, lo que complica la extracción de patrones consistentes [?].
- **Variabilidad intra e interpaciente:** La morfología del tejido puede variar ampliamente, tanto entre diferentes pacientes como dentro de distintas regiones de la misma WSI.

Para hacer viable el procesamiento computacional, es necesario fragmentar la WSI en regiones más pequeñas denominadas *parches* o *tiles*, que pueden ser analizados individualmente por modelos de aprendizaje profundo. Esta **metodología** basada en parches permite reducir la complejidad espacial global y focalizar la extracción de características locales, pero no elimina los desafíos de heterogeneidad ni la necesidad de técnicas que integren contexto global sin etiquetas locales.

Debido a que las anotaciones clínicas suelen disponerse únicamente a nivel de WSI (etiquetas globales), la fragmentación en parches crea un problema de supervisión débil: no se conoce a priori qué parches individuales son informativos para la etiqueta de la imagen completa. Por esta razón, métodos de aprendizaje que prescinden de anotaciones detalladas a nivel de instancia como los **paradigmas** de *Multiple Instance Learning* (MIL), son particularmente adecuados para el análisis de WSIs, ya que permiten entrenar modelos directamente con etiquetas slide-level sin requerir anotaciones de regiones de interés [?] [?].

Finalmente, un pipeline computacional robusto para WSIs debe considerar varias etapas de preprocesamiento, tales como:

- **Normalización de tinción:** Para reducir la variabilidad de color entre diferentes sesiones de escaneo y laboratorios.
- **Filtrado de regiones no informativas:** Para eliminar parches que contienen poco o ningún tejido relevante.
- **Estrategias de muestreo y balance de clases:** Para asegurar una representación adecuada de los patrones clínicamente relevantes durante el entrenamiento del modelo.

Estas decisiones de preprocesamiento y representación son fundamentales para asegurar que los modelos aprendan de manera eficiente a partir de datos con supervisión débil, maximizando su capacidad para generalizar a nuevas WSIs sin requerir anotaciones exhaustivas.

4.4. Redes neuronales convolucionales como extractores de características

Las **redes neuronales convolucionales (CNN)** constituyen una de las familias de modelos más utilizadas en visión por computador y en el campo del análisis de imágenes médicas, dado que permiten aprender automáticamente representaciones jerárquicas de características directamente de los datos de imagen [?, ?]. En particular, las CNN han revolucionado el análisis de imágenes histopatológicas por su capacidad de identificar patrones morfológicos complejos que son relevantes para tareas diagnósticas y subtipificación de tumores.

En el contexto de la histopatología digital, las CNN se emplean principalmente como **extractores de características a nivel de parche** debido a que las imágenes de diapositivas completas (WSIs) son demasiado grandes para ser procesadas directamente por modelos de visión estándar. Convertir cada parche en un **vector de características (embedding)** permite capturar información local relevante, como texturas, estructuras celulares y patrones tisulares, que luego puede utilizarse como entrada para modelos de agregación posteriores [?] [?] [?].

A continuación, se presenta la Figura 4.3 que ilustra la arquitectura típica de una CNN.

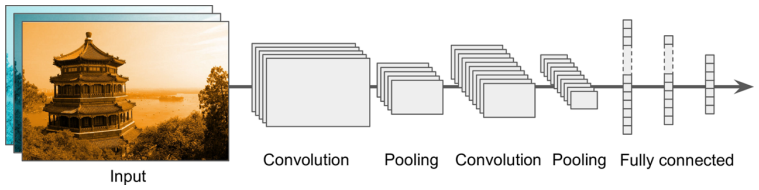

Figura 4.3: Arquitectura típica de una red neuronal convolucional, con capas de convolución, activación y *pooling* que permiten extraer características jerárquicas [?].

Una estrategia comúnmente adoptada en aplicaciones de histopatología digital es el uso de **transfer learning**, donde arquitecturas de CNN pre-entrenadas en conjuntos de datos extensos como ImageNet son adaptadas para tareas específicas de clasificación o extracción de características sobre parches de tejido [?] [?]. Modelos como ResNet, VGG y EfficientNet han demostrado ser eficaces como extractores de características en dominios médicos, reduciendo el tiempo de entrenamiento y mejorando la generalización cuando la cantidad de datos anotados es limitada.

Durante el proceso de entrenamiento de los modelos, se emplean técnicas de **aumentación de datos**, como rotaciones, volteos y variaciones de color, para expandir de forma virtual el conjunto de entrenamiento y mitigar el riesgo de sobreajuste [?]. Además, la normalización de tinción y ajustes de color suelen aplicarse como pasos adicionales de preprocesamiento en histopatología digital para reducir la variación entre diferentes instituciones o dispositivos de escaneo.

En el marco de **Aprendizaje de Instancias Múltiples (MIL)**, los embeddings extraídos por la CNN de cada parche son posteriormente combinados mediante mecanismos de agregación, como el *pooling* o mecanismos de atención, para generar una predicción a nivel de WSI sin requerir anotaciones localizadas exhaustivas [?].

4.5. Aprendizaje de Instancias Múltiples (MIL)

El **Aprendizaje de Instancias Múltiples (MIL)** se presenta como una solución para escenarios en los que las etiquetas están disponibles únicamente a nivel de muestra completa, o **"bag"**. En este contexto, una WSI es interpretada como un *bag* de parches, y la tarea del modelo consiste en inferir la etiqueta global a partir de las características de las instancias locales. Si bien las primeras formulaciones se basaban en estrategias de agregación simples, como el *max pooling*, bajo la premisa de que la presencia de un solo parche positivo determinaba el diagnóstico, este enfoque ignoraba la heterogeneidad del tejido y las cruciales relaciones espaciales entre parches [?] [?].

Para clarificar este planteamiento, la Figura 4.4 muestra un esquema de una WSI dividida en parches (instancias). En MIL las etiquetas se asignan exclusivamente a nivel de *bag* (WSI) y no a cada parche individual; esa es la diferencia clave frente al aprendizaje supervisado a nivel de instancia.

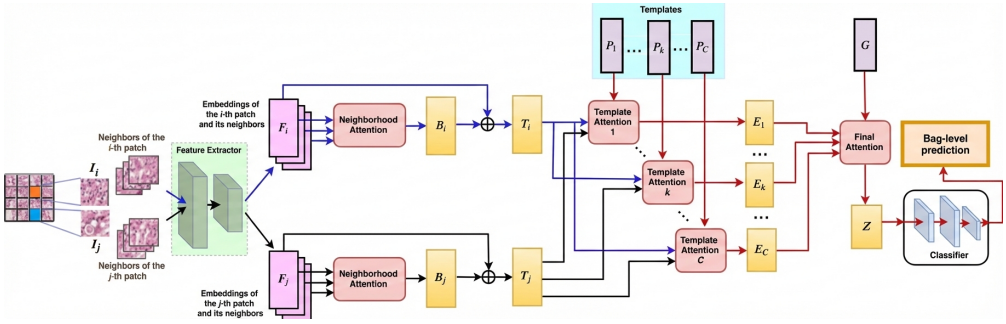

Figura 4.4: Esquema: WSI dividida en parches (instancias). En MIL la etiqueta se proporciona a nivel de bag (WSI) y no a nivel de instancia (parche).

La investigación reciente ha impulsado variantes más sofisticadas del Aprendizaje de Instancias Múltiples que buscan superar las limitaciones de las estrategias de agregación simples. Entre ellas destacan los **mecanismos de atención**, que asignan pesos diferenciados a cada parche según su relevancia para la predicción global [?] [?]; el **pooling jerárquico**, orientado a capturar información en múltiples escalas y niveles de representación [?] [?]; y los **modelos probabilísticos con regularización espacial**, diseñados para modelar la correlación entre parches vecinos y estimar la incertidumbre asociada a la predicción [?] [?].

Estos avances no solo han mejorado la precisión diagnóstica del MIL, sino que también han potenciado su **explicabilidad clínica**, al permitir la identificación visual de las regiones histológicas más relevantes para el diagnóstico final.

4.5.1. MIL con mecanismos de atención e interpretabilidad

Los mecanismos de atención representan una evolución clave dentro del paradigma de Aprendizaje de Instancias Múltiples (MIL), al permitir que el modelo asigne pesos diferenciados a cada instancia en función de su contribución a la predicción global. A diferencia de las estrategias de agregación tradicionales, como el max o mean pooling, los modelos basados en atención aprenden de manera explícita qué regiones del tejido son más informativas para la tarea de clasificación [?] [?] .

Desde el punto de vista metodológico, la atención permite modelar la heterogeneidad intratumoral presente en las imágenes histopatológicas, capturando la contribución relativa de múltiples parches relevantes en lugar de depender de una única instancia dominante. Este enfoque resulta particularmente adecuado en escenarios de supervisión débil, donde únicamente se dispone de etiquetas a nivel de lámina completa (WSI) [?].

Además de mejorar el rendimiento predictivo, los mecanismos de atención aportan una forma de interpretabilidad post-hoc al facilitar la identificación y visualización de las instancias que influyen de manera más significativa en la decisión del modelo. Esta propiedad es especialmente valiosa en el contexto clínico, donde la transparencia, la trazabilidad y la posibilidad de inspeccionar visualmente las regiones relevantes son factores críticos para la confianza y adopción de sistemas de apoyo al diagnóstico [?] [?].

4.5.2. Extensiones recientes de MIL: Transformers y variantes de CLAM

Adicionalmente a los modelos basados en atención descritos previamente, la literatura reciente ha explorado extensiones que integran componentes adicionales como módulos de **transformers** o variantes de agregación más sofisticadas, con el objetivo de capturar relaciones complejas entre las instancias que componen un bag. Estas propuestas buscan modelar de manera más explícita las posibles dependencias entre parches y enriquecer la representación del bag a partir de interacciones más complejas entre las características extraídas de cada instancia.

Un ejemplo prominente de estas aproximaciones es el uso de transformers dentro del marco MIL, donde mecanismos de *self-attention* permiten modelar relaciones globales entre los parches de una WSI sin suponer independencia entre ellos. En el trabajo de Sens *et al.*, se incorpora una pérdida de bag embedding para reforzar la capacidad discriminativa de un transformer aplicado a MIL, demostrando mejoras en conjuntos de datos histopatológicos estandarizados [**?**]. De manera similar, Xiong *et al.* proponen un esquema jerárquico de atención-guía para transformers que explota múltiples resoluciones dentro de una misma WSI, lo cual favorece la identificación de regiones discriminativas de manera más holística [**?**].

Paralelamente, enfoques como CLAM (Clustering-Constrained Attention MIL) han extendido la idea del mecanismo de atención al introducir ramas múltiples de atención y restricciones de agrupamiento para refinar las representaciones de instancias y mejorar la capacidad de clasificación en escenarios con etiquetas de nivel de bag [**?**]. Estas variantes han mostrado rendimiento prometedor en diversas tareas de clasificación, incluida la subtipificación y detección de regiones relevantes en tejidos, al tiempo que generan mapas de atención que pueden utilizarse como evidencia diagnóstica visual.

No obstante, estas arquitecturas más complejas también implican compromisos metodológicos y prácticos importantes. Los modelos basados en transformers suelen requerir conjuntos de datos amplios o estrategias de pre-entrenamiento especializadas para evitar el sobreajuste, dado el elevado número de parámetros inherente a su estructura de *self-attention*. Además, la implementación de múltiples ramas de atención o submódulos interdependientes incrementa la complejidad computacional y la demanda de recursos de memoria durante el entrenamiento y la inferencia.

Dadas las restricciones de disponibilidad de datos, los objetivos de interpretabilidad y la necesidad de una evaluación reproducible en el contexto clínico, el presente trabajo se centra en modelos de atención MIL más simples, tales como ABMIL y SmABMIL, que proporcionan una interpretabilidad post-hoc directa y un balance práctico entre desempeño y complejidad. Esta elección permite una implementación robusta y eficiente sin incurrir en la sobrecarga computacional y de diseño que introducen los transformers o las variantes extendidas de CLAM.

4.6. Métricas de evaluación y análisis

La evaluación de modelos de aprendizaje profundo en histopatología digital requiere un enfoque que trascienda métricas puramente técnicas y considere su impacto clínico real. En escenarios de *supervisión débil*, como aquellos abordados mediante *Multiple Instance Learning* (MIL), las predicciones se realizan a nivel de *bag*, es decir, a nivel de *Whole Slide Image* (WSI), mientras que las instancias individuales (parches) carecen de etiquetas explícitas. Por esta razón, todas las métricas empleadas en este trabajo se definen y analizan exclusivamente a nivel de WSI, en coherencia con la formulación del problema, con el paradigma MIL adoptado y con la práctica clínica real, en la cual las decisiones diagnósticas se toman a nivel de paciente y no

sobre regiones aisladas de tejido [?] [?].

Dado que las tareas de clasificación histopatológica suelen presentar desbalance entre clases y consecuencias clínicas asimétricas, la evaluación del desempeño no puede sustentarse en una única métrica. En su lugar, se adopta un conjunto complementario de métricas que capturan distintos aspectos del comportamiento del modelo a nivel de WSI:

- **Área bajo la curva ROC (AUC-ROC):** La AUC mide la capacidad discriminativa del modelo para ordenar correctamente las WSIs positivas y negativas a lo largo de todos los posibles umbrales de decisión. Esta métrica resulta particularmente adecuada en contextos clínicos, ya que se basa en las probabilidades de predicción generadas a nivel de bag y es independiente del umbral de clasificación, además de ser robusta frente al desbalance de clases [?].
- **Precisión (Precision):** Define la proporción de WSIs clasificadas como positivas que son efectivamente positivas. En aplicaciones clínicas, una alta precisión reduce la incidencia de falsos positivos, los cuales pueden derivar en procedimientos invasivos innecesarios, ansiedad en el paciente y sobrecarga del sistema de salud [?].
- **Sensibilidad (Recall):** También conocida como tasa de verdaderos positivos, cuantifica la proporción de WSIs positivas correctamente identificadas por el modelo. Desde una perspectiva clínica, la sensibilidad es una métrica crítica, ya que los falsos negativos representan el riesgo de omitir un diagnóstico relevante, con potencial impacto directo en el pronóstico del paciente [?].
- **F1-score:** Corresponde a la media armónica entre precisión y sensibilidad, proporcionando una medida balanceada del desempeño del modelo. Esta métrica es especialmente útil en escenarios con clases desbalanceadas, ya que penaliza simultáneamente tanto los falsos positivos como los falsos negativos, ofreciendo una estimación más estable y representativa del rendimiento global del sistema en escenarios de desbalance de clases [?].
- **Coeficiente de concordancia de Cohen (Kappa):** El coeficiente Kappa cuantifica el grado de acuerdo entre las predicciones del modelo y las etiquetas reales de las WSIs, corrigiendo el acuerdo esperado por azar. A diferencia de métricas como la exactitud, Kappa permite evaluar el desempeño del modelo en relación con el nivel de concordancia que típicamente se observa entre observadores humanos, lo cual resulta particularmente relevante en histopatología, donde existe variabilidad interobservador incluso entre patólogos expertos [?].

Más allá de las métricas cuantitativas tradicionales, la **interpretabilidad** constituye un requisito fundamental para la adopción clínica de modelos basados en MIL. En este contexto, los mecanismos de atención permiten identificar qué subconjuntos de parches dentro de una WSI contribuyen de manera más significativa a la predicción final, sin requerir anotaciones explícitas a nivel de parche, lo cual es coherente con el esquema de supervisión débil propio del aprendizaje MIL [?] [?]. Esta capacidad interpretativa no solo incrementa la confianza en el sistema, sino que también alinea el proceso de inferencia del modelo con el razonamiento diagnóstico del patólogo.

Adicionalmente, la evaluación del modelo se complementa con el análisis conceptual de la **incertidumbre en la predicción**, lo cual resulta crucial en aplicaciones clínicas de alto impacto. La identificación de casos con baja confianza permite establecer mecanismos de derivación a revisión manual, reforzando la seguridad del sistema y promoviendo un uso responsable de modelos de aprendizaje automático en entornos médicos [?].

En conjunto, este esquema de evaluación garantiza que el desempeño del modelo sea analizado no solo desde una perspectiva computacional, sino también desde su relevancia clínica, su coherencia con el paradigma MIL y su alineación con prácticas diagnósticas reales en histopatología digital.

4.7. Estado del Arte

La investigación en aprendizaje débilmente supervisado sobre Whole Slide Images (WSI) ha avanzado rápidamente en los últimos años, motivada por la necesidad de reducir la carga de anotación experta y por el potencial clínico de modelos reproducibles y explicables. Varios trabajos han propuesto arquitecturas y estrategias que permiten agregar evidencia local (parches) en predicciones a nivel de muestra (bag), y han explorado alternativas para incorporar contexto espacial, medidas de incertidumbre y visualizaciones de interpretabilidad.

Para contextualizar la presente investigación, a continuación se revisan los antecedentes más relevantes que abordan estas problemáticas. Dichos estudios justifican las decisiones metodológicas de este proyecto al ilustrar las principales tendencias y desafíos en el campo, sirviendo como base teórica para el diseño de la solución propuesta:

4.7.1. CAMIL: channel attention-based multiple instance learning for whole slide image classification

En este trabajo, los autores proponen una arquitectura denominada FA-MILNet (Feature Aggregation Multiple Instance Learning Network), orientada a abordar los desafíos del aprendizaje débilmente supervisado en la clasificación de Whole Slide Images (WSI). La propuesta surge como respuesta a las limitaciones de los enfoques MIL tradicionales, los cuales suelen perder información discriminativa al emplear estrategias de agregación simples como el max pooling y no modelan adecuadamente la heterogeneidad intralaminar.

FA-MILNet introduce un esquema de agregación jerárquica asistida por atención, integrando información local y global mediante tres componentes principales: extracción de características basada en CNN, un módulo de atención para ponderar la relevancia de los parches y una rama dual de agregación. Esta arquitectura permite mejorar simultáneamente el desempeño predictivo y la interpretabilidad del modelo [?].

Evaluado en conjuntos de datos de cáncer de colon, mama y ganglios linfáticos (Camelyon16, BACH y CRC), el modelo supera enfoques como ABMIL, CLAM y DSMIL en métricas como AUC y F1-score. Si bien el estudio no se enfoca en cáncer de próstata, su diseño modular y su capacidad para operar sin anotaciones patch-level lo convierten en un antecedente altamente pertinente para este proyecto, al validar el uso de mecanismos de atención como estrategia central para la agregación de información en escenarios de supervisión débil [?].

4.7.2. A Probabilistic MIL Model with Spatial Regularization for Weakly Supervised Histopathology Image Analysis

Este trabajo presenta un modelo de Aprendizaje de Instancias Múltiples de carácter probabilístico, basado en procesos gaussianos e integrado con un término de regularización espacial inspirado en el modelo de Ising. La propuesta parte de la hipótesis de que los parches histológicamente cercanos presentan una alta correlación diagnóstica, por lo que modelar explícitamente estas dependencias espaciales puede mejorar la

robustez del aprendizaje bajo supervisión débil [?].

A diferencia de los modelos MIL determinísticos, esta aproximación permite estimar distribuciones de probabilidad sobre las etiquetas de las instancias, proporcionando una cuantificación explícita de la incertidumbre asociada a las predicciones. Esta característica resulta particularmente relevante en contextos clínicos, donde la identificación de casos ambiguos y la evaluación de la confianza del modelo son aspectos críticos para su adopción práctica.

4.7.3. Graph-based Multiple Instance Learning for Whole Slide Image Classification

En este estudio se propone una variante de MIL basada en grafos, donde cada parche de una WSI se modela como un nodo y las relaciones espaciales entre parches se representan mediante aristas en un grafo no dirigido. Sobre esta estructura se emplea una Graph Neural Network (GNN), permitiendo la propagación de información contextual entre regiones del tejido y facilitando la captura de patrones arquitectónicos globales [?].

Este enfoque ofrece ventajas claras en términos de interpretabilidad estructural y modelado explícito del contexto espacial. Sin embargo, introduce una complejidad computacional significativa asociada tanto a la construcción del grafo como al entrenamiento del modelo, además de requerir decisiones de diseño adicionales sobre la conectividad y los criterios de vecindad.

Dichas consideraciones exceden el alcance experimental del presente proyecto, lo que motivó la selección de arquitecturas más ligeras basadas en mecanismos de atención directa, que permiten un análisis interpretativo más sencillo y una evaluación reproducible bajo restricciones computacionales realistas [?].

4.7.4. Deep Recurrent Attention MIL for Histopathological Image Analysis

El enfoque planteado en este artículo se distingue por combinar mecanismos de atención con redes neuronales recurrentes (RNN) dentro del marco de MIL. En lugar de tratar los parches como elementos independientes, los autores introducen una secuencia ordenada basada en la posición espacial de cada uno, la cual es procesada por una RNN con atención. Este diseño permite capturar no solo las características locales, sino también las relaciones de largo alcance y las dependencias contextuales más profundas a lo largo de toda la WSI [?].

Gracias a esta arquitectura secuencial, el modelo demuestra un mejor rendimiento en la identificación de estructuras glandulares complejas, un desafío particular en escenarios con gran variabilidad morfológica entre pacientes. Además, el mecanismo de atención adaptativa complementa este diseño, facilitando interpretaciones más precisas del modelo al resaltar de forma dinámica las regiones más relevantes. La principal lección de este trabajo es que la secuencia y el orden de los parches, y no solo su presencia, pueden aportar información crucial. A diferencia de los enfoques basados en grafos que modelan la vecindad de forma explícita, este método explora una representación secuencial, lo que invita a considerar cómo la estructura de los datos de entrada puede ser modificada para extraer de manera más efectiva la información contextual [?].

4.7.5. Hierarchical Pooling in MIL for Histopathological Subtype Classification

Este trabajo propone un esquema de agregación jerárquica dentro del paradigma MIL, en el cual los parches primero se agrupan en regiones intermedias según su proximidad o similitud morfológica. Posteriormente, se aplica un segundo nivel de pooling que combina la información de estas regiones para obtener la predicción final del bag. Esta estructura multinivel permite capturar patrones de organización tisular más amplios, que son clave en el diagnóstico de ciertos subtipos de cáncer. La metodología incorpora tanto técnicas de agrupamiento automático como atención regional, y ha mostrado un desempeño superior en conjuntos de datos con alta heterogeneidad intratumoral, al mantener el balance entre granularidad local y contexto estructural [?].

Este antecedente es relevante porque, al igual que FA-MILNet, enfatiza la importancia de una agregación multinivel. Sin embargo, se distingue por su énfasis en la creación de "grupos" lógicos de parches antes de la agregación final, lo que es una estrategia diferente a la simple ponderación por atención. Por otro lado. la idea de agrupar parches basándose en la morfología o la proximidad podría ser una técnica poderosa a considerar para proyectos que buscan capturar la complejidad de la heterogeneidad tumoral [?].

4.7.6. Síntesis crítica y posicionamiento del presente trabajo

La revisión del estado del arte evidencia que el Aprendizaje de Instancias Múltiples (MIL) aplicado a Whole Slide Images ha experimentado una evolución sustancial, incorporando mecanismos de atención, modelado espacial explícito y esquemas de agregación jerárquica. Estas aproximaciones han demostrado mejoras significativas tanto en desempeño predictivo como en capacidad interpretativa, particularmente en escenarios caracterizados por supervisión débil y elevada heterogeneidad morfológica.

No obstante, muchos de estos métodos de última generación, tales como CLAM o arquitecturas MIL basadas en Transformers, dependen de componentes adicionales como pseudo-etiquetado, clustering previo, modelado explícito de grafos o mecanismos de autoatención de alta complejidad. Si bien estas estrategias pueden mejorar el rendimiento en determinados contextos, también introducen un incremento considerable en el costo computacional, la sensibilidad a hiperparámetros y la dificultad de interpretación directa, factores que limitan su aplicabilidad práctica y su evaluación reproducible en entornos clínicos reales.

Bajo esta premisa, el presente trabajo se posiciona deliberadamente en un punto de equilibrio entre capacidad predictiva, interpretabilidad y viabilidad experimental. En particular, se adopta un marco de MIL basado en mecanismos de atención explícitos (ABMIL y SmABMIL), los cuales permiten ponderar la contribución relativa de cada parche a la predicción global sin requerir anotaciones a nivel de instancia ni imponer estructuras espaciales adicionales. Esta elección metodológica se alinea con las restricciones reales de disponibilidad de datos en histopatología digital y facilita un análisis interpretativo claro y directamente asociado al proceso de decisión del modelo.

Asimismo, este enfoque establece una base sólida y extensible para trabajos futuros, en los cuales podrían incorporarse modelos jerárquicos, probabilísticos o basados en grafos una vez se disponga de mayores recursos computacionales o anotaciones complementarias, preservando siempre la coherencia clínica y la trazabilidad interpretativa del sistema propuesto.

Capítulo 5

Definición de Requisitos

La definición de requisitos constituye una etapa fundamental en el desarrollo del presente proyecto, ya que permite establecer de manera explícita las capacidades, restricciones y criterios de validación del sistema propuesto. Estos requisitos se derivan directamente del análisis del problema, los objetivos planteados y los antecedentes revisados en el estado del arte, funcionando como un marco de referencia para evaluar la coherencia y el alcance de la solución desarrollada.

Dado que el objetivo central de esta investigación es el diseño y evaluación de un modelo de MIL para el análisis de imágenes histopatológicas digitales de próstata bajo un esquema de supervisión débil, los requisitos se formulan desde una perspectiva metodológica y científica, y sirven como puente entre la fase de concepción del sistema y su posterior implementación y evaluación experimental.

5.1. Alcance del sistema propuesto

El sistema desarrollado en este proyecto tiene como finalidad analizar *Whole Slide Images* (WSI) de biopsias de próstata mediante un enfoque de aprendizaje débilmente supervisado, con el objetivo de inferir etiquetas diagnósticas a nivel de muestra completa. En este contexto, cada WSI es modelada como un conjunto de parches (instancias), a partir de los cuales se extraen representaciones profundas que posteriormente son agregadas mediante un modelo MIL con mecanismos de atención.

El alcance del sistema se limita al análisis computacional de imágenes histopatológicas previamente digitalizadas y no contempla procesos clínicos como la adquisición de muestras, la digitalización de las láminas ni la validación clínica prospectiva de los resultados. Asimismo, el sistema está diseñado como una herramienta de apoyo a la investigación y no como un producto clínico listo para su despliegue en entornos hospitalarios.

5.2. Requisitos funcionales

Desde el punto de vista funcional, el sistema debe satisfacer las siguientes capacidades fundamentales:

- **Procesamiento de imágenes de gran escala:** El sistema debe permitir el análisis eficiente de WSIs de alta resolución, descomponiéndolas en parches que capturen información morfológica relevante sin requerir anotaciones a nivel de instancia.
- **Pipeline de preprocesamiento:** Se requiere una etapa de preprocesamiento orientada a la generación de parches homogéneos y filtrados, eliminando regiones no informativas como fondo, ruido o artefactos de preparación.
- **Extracción de características:** El sistema debe incorporar un mecanismo basado en redes neuronales convolucionales (CNN) preentrenadas para obtener representaciones vectoriales (*embeddings*)

discriminativas de cada parche. Estas representaciones constituyen la entrada principal del modelo MIL encargado de inferir la predicción a nivel de WSI.

- **Mecanismos de atención:** Es necesaria la integración de mecanismos de atención que asignen pesos diferenciales a las instancias en función de su relevancia diagnóstica, permitiendo identificar las regiones del tejido que contribuyen de manera dominante a la decisión del modelo.
- **Generación de métricas y visualizaciones:** El sistema debe generar métricas de evaluación cuantitativas a nivel de lámina completa (slide), así como visualizaciones que reflejen la distribución de la atención sobre la WSI, facilitando el análisis posterior de los resultados.

5.3. Requisitos no funcionales

En términos de atributos de calidad y operatividad, el sistema debe garantizar:

- **Reproducibilidad:** Los experimentos deben ser replicables mediante el uso de semillas aleatorias controladas, configuraciones explícitas y una organización estructurada del código y los experimentos.
- **Escalabilidad:** El sistema debe poder manejar WSIs con un número elevado de parches, sin que ello implique un crecimiento prohibitivo en los tiempos de cómputo, mediante un diseño modular que desacople la extracción de características de la agregación MIL.
- **Interpretabilidad:** Dado el dominio clínico del problema, el sistema debe proporcionar mecanismos que permitan analizar y justificar las decisiones del modelo, evitando enfoques de tipo "caja negra". Si bien no se busca una interpretabilidad causal estricta, sí se requiere una interpretabilidad post-hoc basada en mecanismos de atención y visualización de regiones relevantes.
- **Compatibilidad:** El sistema debe ser compatible con infraestructuras estándar en investigación de ciencia de datos, como entornos basados en GPU y *frameworks* de aprendizaje profundo ampliamente adoptados.

5.4. Restricciones del proyecto

El desarrollo del sistema se encuentra condicionado por diversas restricciones inherentes tanto al dominio clínico como al contexto académico de la investigación. Estas limitaciones se describen a continuación:

1. **Supervisión débil:** La disponibilidad de etiquetas se limita al nivel de WSI, lo que impide el uso de enfoques supervisados a nivel de parche.
2. **Heterogeneidad morfológica:** La alta variabilidad intratumoral en las biopsias de próstata dificulta la agregación simple de información y refuerza la necesidad de mecanismos de atención y modelos capaces de manejar dicha variabilidad.
3. **Recursos computacionales:** El tamaño de las WSIs y la cantidad de parches generados imponen limitaciones en términos de memoria y tiempo de entrenamiento, lo que obliga a adoptar estrategias de muestreo, reducción de dimensionalidad o separación de etapas dentro del pipeline.
4. **Validación clínica:** Al tratarse una investigación académica, el proyecto está limitado en términos de validación clínica directa, por lo que los resultados deben ser interpretados como evidencia experimental y no como conclusiones clínicas definitivas.

5.5. Criterios de validación

Los criterios para validar el sistema se definen bajo los siguientes parámetros:

- **Evaluación cuantitativa:** El desempeño del modelo debe ser evaluado mediante métricas cuantitativas apropiadas para problemas de clasificación a nivel de WSI, tales como accuracy, AUC y F1-score, permitiendo una comparación objetiva entre diferentes configuraciones del modelo.
- **Estabilidad y generalización:** Se debe evaluar la consistencia del modelo frente a distintas particiones de los datos (validación cruzada) con el fin de analizar su capacidad de generalización. Este criterio es especialmente relevante dado el tamaño limitado de los conjuntos de datos histopatológicos y la alta variabilidad entre muestras.
- **Coherencia histopatológica:** Como criterio cualitativo, los mapas de atención generados deben ser coherentes desde un punto de vista histopatológico, es decir, que las regiones destacadas por el modelo correspondan a áreas con características morfológicas plausibles según el conocimiento experto. Este análisis no pretende sustituir la evaluación clínica, pero sí aportar evidencia sobre la consistencia y utilidad interpretativa del enfoque propuesto.

Capítulo 6

Diseño del Pipeline

6.1. Diseño del Pipeline Experimental

Este capítulo presenta el diseño conceptual y técnico del pipeline experimental propuesto para el análisis de imágenes histopatológicas de próstata mediante el paradigma de Aprendizaje de Instancias Múltiples (MIL). Su propósito es describir de forma estructurada cómo se concibió la solución, qué componentes la conforman y cómo se organiza el flujo de datos desde la adquisición de las Whole Slide Images (WSI) hasta la generación de predicciones a nivel de lámina, sin anticipar resultados cuantitativos ni interpretaciones empíricas.

El diseño del pipeline fue desarrollado siguiendo un enfoque incremental basado en sprints, lo que permitió descomponer el problema en etapas claramente delimitadas, garantizar trazabilidad entre decisiones técnicas y facilitar la alineación entre el diseño metodológico y la implementación computacional realizada posteriormente en los notebooks experimentales.

Cada sprint representa una unidad lógica de diseño, estableciendo responsabilidades precisas dentro del sistema y definiendo explícitamente qué aspectos quedan fijados en esta etapa y cuáles serán abordados operativamente en el capítulo de Metodología Experimental y Resultados.

6.1.1. Visión General del Pipeline Metodológico

El pipeline completo fue representado gráficamente con el fin de proporcionar una visión global del flujo de procesamiento y de las interacciones entre sus distintos componentes. Debido a su complejidad y nivel de detalle, el diagrama fue dividido en dos partes complementarias que, en conjunto, describen el proceso completo desde los datos crudos hasta la salida del modelo.

La Figura 6.1 ilustra la primera parte del pipeline, abarcando la adquisición de las WSI, su partición en parches, el preprocesamiento inicial y la extracción de características a nivel de instancia.

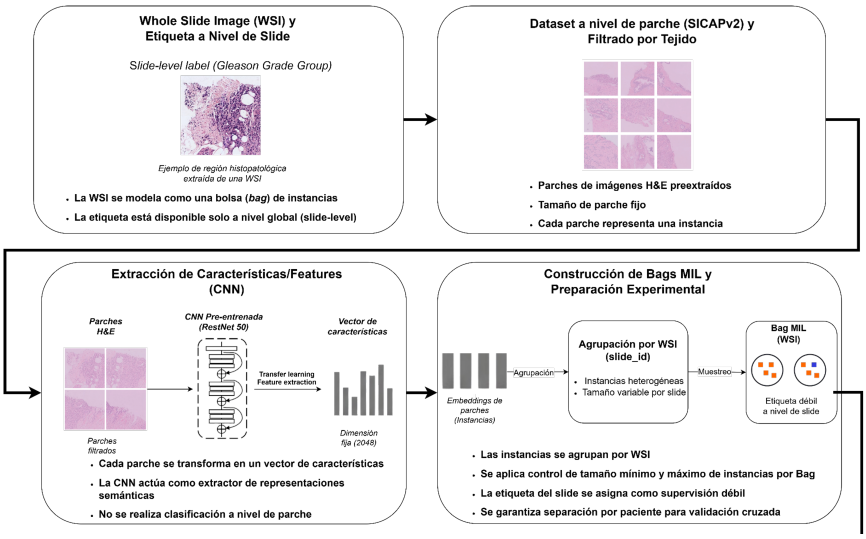

Figura 6.1: Primera parte del pipeline propuesto: adquisición de WSI, partición en patches, preprocesamiento y extracción de características a nivel de instancia.

Por otro lado, la Figura 6.2 presenta la segunda parte del pipeline, correspondiente a la organización de instancias en bolsas MIL, el mecanismo de atención, la agregación de información y la generación de predicciones a nivel de lámina.

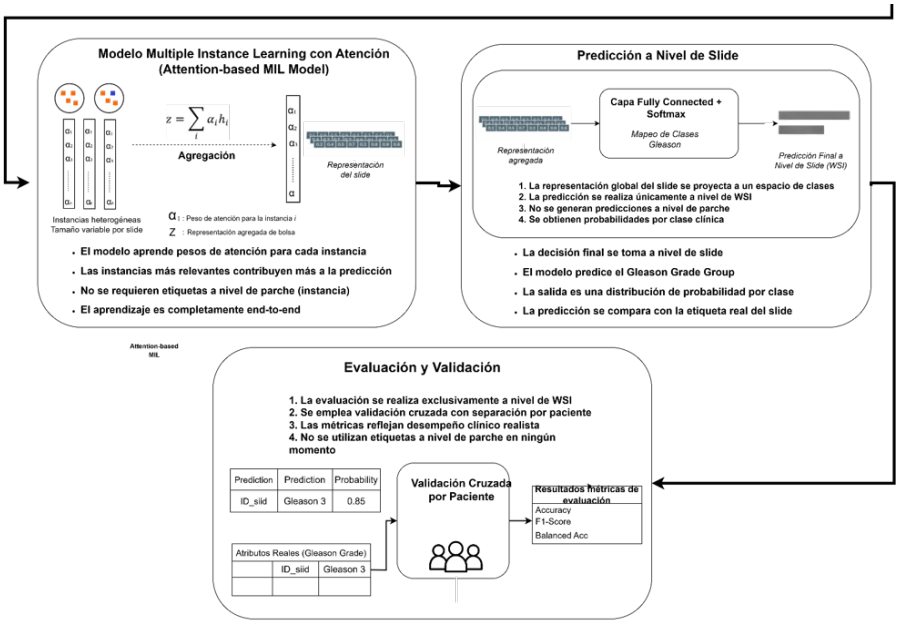

Figura 6.2: Segunda parte del pipeline: modelo MIL con mecanismo de atención, agregación de instancias, predicción y evaluación conceptual.

Ambos diagramas deben interpretarse como un único flujo continuo de procesamiento, en el cual cada etapa transforma progresivamente la información histopatológica desde su representación cruda hasta una decisión clínica modelada.

6.1.2. Sprint 0: Infraestructura, Entorno y Reproducibilidad

El Sprint 0 tuvo como objetivo establecer una infraestructura computacional reproducible y adecuada para el procesamiento de imágenes histopatológicas a gran escala.

Se definió un entorno de ejecución basado en GPU utilizando Google Colaboratory, seleccionando explícitamente bibliotecas y versiones compatibles con patología digital y aprendizaje profundo, tales como PyTorch, NumPy, Pandas y OpenSlide.

Asimismo, se diseñó una estructura de directorios estandarizada que separa datos crudos, datos procesados, artefactos intermedios y salidas experimentales, garantizando trazabilidad completa del flujo de datos y facilitando la replicación de los experimentos. Cabe resaltar que en este sprint no se ejecutan experimentos ni se generan métricas, sino que se fijan las bases técnicas sobre las cuales se construye todo el pipeline experimental.

6.1.3. Sprint 1: Comprensión del Dataset y Diseño de Particiones

El Sprint 1 estuvo orientado al análisis estructural del dataset SICAPv2 y al diseño de una estrategia de partición experimental clínicamente válida.

Se realizó una revisión exhaustiva de la estructura del conjunto de datos, verificando la correspondencia entre parches, anotaciones clínicas y metadatos asociados, así como la relación jerárquica entre pacientes, láminas y parches.

Un aspecto central de este sprint fue el diseño de una estrategia de validación que evitara explícitamente la fuga de datos (*data leakage*), definiendo particiones basadas en identificadores de paciente como unidad indivisible.Este diseño garantiza que ninguna información de un mismo paciente esté presente simultáneamente en conjuntos de entrenamiento y validación, alineando la evaluación con escenarios clínicos realistas.

6.1.4. Sprint 2: Construcción del *Dataset Manifest*

El Sprint 2 definió el mecanismo central de control del flujo de datos mediante la construcción del archivo `dataset_manifest.csv`, concebido como la fuente única de verdad del pipeline experimental.

El manifiesto fue diseñado para consolidar toda la información relevante de cada parche, incluyendo identificadores de paciente y lámina, asignación a fold, conjunto experimental, etiqueta clínica global y rutas a los archivos correspondientes.

Desde el punto de vista metodológico, este archivo actúa como un contrato formal entre las distintas etapas del pipeline, asegurando consistencia, trazabilidad y control total de las particiones experimentales. Las validaciones estructurales asociadas a este manifiesto fueron definidas conceptualmente en este sprint, dejando su ejecución concreta para la metodología experimental.

6.1.5. Sprint 3: Ingeniería de Características y Construcción de Bolsas MIL

El Sprint 3 se enfocó en el diseño del proceso de ingeniería de características y en la definición formal de las bolsas de instancias requeridas por el paradigma MIL.

Se estableció el uso de un extractor de características basado en una red neuronal convolucional profunda preentrenada, con el objetivo de transformar cada parche histopatológico en una representación vectorial compacta y semánticamente rica.

Previamente, se definieron criterios de filtrado visual para descartar regiones no informativas, reduciendo ruido y optimizando el uso de recursos computacionales.A nivel conceptual, se definió que cada bolsa MIL corresponde a una WSI completa, agrupando un número variable de instancias, lo cual refleja la heterogeneidad inherente al tejido prostático.

6.1.6. Sprint 4: Diseño de Modelos MIL

El Sprint 4 abordó el diseño de las arquitecturas de Aprendizaje de Instancias Múltiples empleadas en el sistema.

Se definieron múltiples estrategias de agregación a nivel de bolsa, incluyendo enfoques basados en promedios, máximos y mecanismos de atención, todas concebidas para operar sobre las mismas representaciones y particiones.

Desde el punto de vista de diseño, se estableció que el entrenamiento y la evaluación de los modelos se realicen de forma independiente por fold, respetando estrictamente las particiones definidas en el manifiesto. Este sprint describe el espacio de modelos considerado, sin introducir comparaciones empíricas ni resultados cuantitativos.

6.1.7. Sprint 5: Interpretabilidad y Análisis Conceptual

El Sprint 5 definió el componente de interpretabilidad del pipeline, concebido como un requisito fundamental para aplicaciones clínicas.

En particular, se diseñó el uso de mecanismos de atención como herramienta para asignar pesos de importancia a las instancias dentro de cada bolsa, permitiendo identificar regiones histopatológicas relevantes desde un punto de vista conceptual.

Asimismo, se estableció que el análisis del comportamiento del sistema incluya herramientas visuales y métricas agregadas que permitan evaluar estabilidad y coherencia, dejando su cálculo y análisis detallado para capítulos posteriores.

6.1.8. Resumen del Diseño del Pipeline

En conjunto, el pipeline diseñado establece un flujo de trabajo modular, reproducible y clínicamente alineado para el análisis de imágenes histopatológicas mediante MIL.

La separación explícita por sprints permite distinguir con claridad entre decisiones de diseño y ejecución experimental, minimizando ambigüedades metodológicas y preparando una transición natural hacia el capítulo de Metodología Experimental y Resultados.

Capítulo 7

Metodología Experimental y Resultados

7.1. Metodología Experimental

La presente metodología experimental describe de manera detallada el procedimiento seguido para implementar, ejecutar y validar el pipeline propuesto para el análisis de imágenes histopatológicas prostáticas bajo el paradigma de Aprendizaje de Instancias Múltiples (MIL).

Esta sección se enfoca exclusivamente en la ejecución operativa del diseño metodológico previamente definido, estableciendo con precisión cómo se procesaron los datos, cómo se construyeron las estructuras experimentales y bajo qué condiciones se entrenaron los modelos, sin anticipar resultados cuantitativos ni interpretaciones clínicas.

Todas las decisiones metodológicas descritas en este capítulo se encuentran directamente respaldadas por la implementación realizada en los notebooks experimentales, garantizando coherencia, trazabilidad y reproducibilidad del proceso completo.

7.1.1. Configuración Experimental y Entorno de Ejecución

Los experimentos fueron desarrollados en un entorno de computación acelerada por GPU, utilizando la plataforma Google Colaboratory como infraestructura principal de ejecución.

El entrenamiento y la inferencia de los modelos se realizaron mediante la librería PyTorch, aprovechando capacidades de cómputo CUDA cuando estuvieron disponibles, lo cual permitió manejar de forma eficiente grandes volúmenes de datos y modelos con un número elevado de parámetros. Por otro lado, para garantizar la reproducibilidad de los experimentos, se fijaron semillas aleatorias para los generadores de números pseudoaleatorios de Python, NumPy y PyTorch, incluyendo tanto ejecuciones en CPU como en GPU.

Adicionalmente, la gestión explícita de dependencias, rutas de archivos y organización del espacio de trabajo fue implementada directamente en el código, asegurando que cada experimento pudiera ser replicado de manera determinística bajo las mismas condiciones computacionales.

7.1.2. Conjunto de Datos y Definición Operativa del Problema

El conjunto de datos utilizado en este estudio corresponde a la base pública SICAPv2, la cual contiene imágenes histopatológicas prostáticas digitalizadas en formato *Whole Slide Image* (WSI), teñidas con Hematoxilina y Eosina (H&E), junto con anotaciones clínicas y máscaras derivadas de la gradación de Gleason [?].

El dataset está compuesto por múltiples láminas provenientes de distintos pacientes, cada una asociada a un puntaje de Gleason primario y secundario, a partir del cual se deriva el grado ISUP correspondiente.

En su formulación original, SICAPv2 fue diseñado para tareas de segmentación y clasificación multiclase de patrones histopatológicos, incluyendo regiones no cancerosas (NC) y patrones Gleason 3, 4 y 5.

Desde el punto de vista clínico, el cáncer de próstata presenta una marcada heterogeneidad espacial, donde múltiples patrones histológicos pueden coexistir dentro de una misma lámina. Esta característica se refleja en la anotación multiclase original del dataset. No obstante, aunque se dispone de información local a nivel de parche sobre patrones histológicos específicos (NC, G3, G4, G5 y variantes cribiformes), dichas etiquetas no fueron utilizadas como señal supervisada durante el entrenamiento de los modelos. En el presente trabajo, la tarea fue formulada como un problema de clasificación binaria a nivel de WSI, distinguiendo entre láminas con presencia de malignidad clínicamente significativa y láminas sin evidencia predominante de tumor agresivo. Esta reformulación responde a dos criterios principales:

- **Criterio clínico:** En escenarios de apoyo diagnóstico inicial, la decisión primaria se centra en determinar la presencia o ausencia de enfermedad relevante, antes de abordar una gradación histológica detallada.

- **Criterio metodológico:** La formulación binaria permite evaluar de manera más controlada la capacidad discriminativa de los modelos bajo supervisión débil, evitando el efecto amplificador del desbalance de clases propio de la clasificación multiclase.

Dado que no se empleó supervisión exhaustiva a nivel de parche para la tarea de clasificación final, el problema se enmarca dentro del paradigma de Aprendizaje de Instancias Múltiples (MIL). Bajo esta formulación, cada WSI se define como una bolsa (*bag*) compuesta por múltiples instancias (*patches*), y la etiqueta global del bag corresponde al diagnóstico clínico asociado a la lámina completa.

Todas las predicciones, métricas y evaluaciones se realizan estrictamente a nivel de bolsa (WSI-level), garantizando coherencia con la formulación de supervisión débil y con el flujo de trabajo clínico real, donde la decisión diagnóstica se emite sobre la lámina completa. Esta definición operativa fundamenta directamente la interpretación de las métricas reportadas en el Capítulo de Resultados, donde el desempeño de las arquitecturas MIL es evaluado exclusivamente a nivel de WSI. En la Figura 7.1 se presentan ejemplos representativos de láminas pertenecientes al dataset SICAPv2, ilustrando la variabilidad morfológica y cromática característica de este tipo de imágenes histopatológicas.

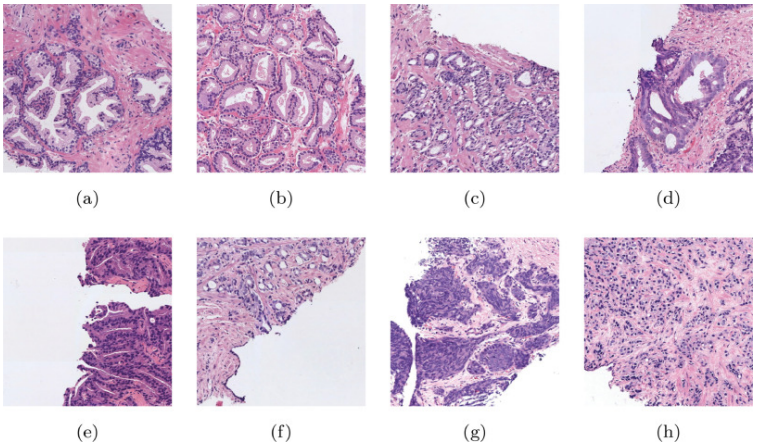

Figura 7.1: Ejemplos representativos de Whole Slide Images del dataset SICAPv2. Se observa la variabilidad morfológica y estructural del tejido prostático teñido con H&E.

7.1.3. Preparación Inicial del Dataset y Control de Integridad

1. **Descarga y organización:** Se realizó la descarga, descompresión y organización estructurada del dataset SICAPv2, preservando la trazabilidad completa desde el origen de los datos hasta su disposición final en el entorno de trabajo.

2. **Verificación de estructura:** Se verificó la estructura real del conjunto de datos en disco, confirmando la existencia y correspondencia entre imágenes, máscaras y archivos de partición oficiales proporcionados por el dataset.

3. **Validación de consistencia:** Durante esta etapa se validó la integridad del par imagen–máscara, la ausencia de archivos corruptos y la coherencia de los identificadores de lámina utilizados a lo largo de las distintas particiones experimentales.

4. **Consolidación de la base de datos:** Este proceso permitió establecer una base de datos limpia, íntegra y estructuralmente consistente antes de iniciar cualquier fase de transformación o entrenamiento posterior.

7.1.4. Estrategia de Partición y Prevención de Fuga de Datos

El dataset SICAPv2 proporciona particiones oficiales para validación cruzada, las cuales fueron empleadas directamente en este trabajo con el fin de garantizar comparabilidad y rigor metodológico. Cada partición fue analizada para confirmar que no existiera solapamiento de información entre los conjuntos de entrenamiento y validación a nivel de paciente o lámina, evitando explícitamente escenarios de fuga de datos (*data leakage*).

Por otro lado, la unidad mínima de partición se definió a nivel de WSI, asegurando que todos los parches derivados de una misma lámina permanezcan agrupados dentro del mismo subconjunto experimental. Este esquema de validación emula de manera más fiel el flujo de trabajo clínico real y permite evaluar la capacidad de generalización del modelo bajo un escenario estrictamente paciente-independiente, donde ningún individuo contribuye simultáneamente a los conjuntos de entrenamiento y validación.

7.1.5. Construcción del *Dataset Manifest*

Con el objetivo de centralizar y controlar el flujo de datos del experimento, se construyó un archivo denominado `dataset_manifest.csv`, concebido como la fuente única de verdad del sistema. En este manifiesto, cada fila corresponde a un parche histopatológico y consolida información clave como su ruta de acceso, la WSI de origen, el fold de validación, el subconjunto experimental, las etiquetas clínicas globales y metadatos asociados.

Para asegurar la coherencia diagnóstica, el grado ISUP se determinó algorítmicamente a nivel de lámina (WSI) basándose en los puntajes de Gleason primario y secundario. Posteriormente, este valor se propagó de forma consistente a todas las instancias (parches) derivadas de dicha lámina.

La integridad del manifiesto se verificó mediante un protocolo de validación diseñado para detectar valores nulos, confirmar la existencia física de los archivos en disco y asegurar una separación estricta entre subconjuntos experimentales, evitando así cualquier sesgo de selección o solapamiento de datos.

7.1.5.1. Definición Formal y Rol Metodológico del Dataset Manifest

El archivo `dataset_manifest.csv` constituye el componente metodológico central del *pipeline* experimental, actuando como el **registro maestro de datos** para la gestión, trazabilidad y control de la información utilizada en todas las fases del estudio. Su función trasciende la mera organización de archivos, estableciéndose como el mecanismo de gobernanza que garantiza la integridad y la reproducibilidad de los experimentos.

Desde una perspectiva formal, el manifiesto define una correspondencia unívoca entre cada instancia (parche histopatológico) y su contexto experimental. Esta estructura encapsula de manera atómica la información crítica: rutas de acceso indexadas, pertenencia a la *Whole Slide Image* (WSI) de origen, identidad del paciente y asignación a particiones específicas. Este diseño permite un desacoplamiento estricto entre la curaduría del *dataset* y la implementación de los modelos de aprendizaje, eliminando dependencias implícitas y asegurando que la lógica de datos sea independiente de los *scripts* de entrenamiento.

Bajo esta configuración, cada fila del manifiesto representa exactamente un parche extraído de una lámina, estableciendo explícitamente su jerarquía dentro del conjunto de datos. En el ámbito experimental, el manifiesto funciona como un contrato formal entre las etapas del flujo de trabajo, asegurando que cualquier proceso posterior —ya sea el entrenamiento de modelos MIL o la validación cruzada— utilice definiciones de bolsas y etiquetas clínicas idénticas. Asimismo, facilita la auditoría técnica y clínica, permitiendo verificar de forma explícita la procedencia de cada instancia y su contribución a las predicciones globales a nivel de lámina.

7.1.5.2. Estructura del Archivo `dataset_manifest.csv`

La estructura del archivo `dataset_manifest.csv` fue diseñada para capturar de manera explícita la información jerárquica inherente a los datos histopatológicos, abarcando niveles de paciente, lámina e instancia. En el cuadro 7.1 se presenta la descripción formal de las columnas que componen el manifiesto, junto con su rol dentro del pipeline experimental.

Cuadro 7.1: Estructura del archivo `dataset_manifest.csv` utilizado como fuente única de verdad del pipeline experimental.

Columna	Descripción
imageName	Nombre del archivo correspondiente al parche histopatológico.
imagePath	Ruta completa al archivo de imagen del parche.
maskPath	Ruta a la máscara asociada al parche, en caso de existir.
maskExists	Indicador booleano que señala la existencia o ausencia de máscara.
wsiId	Identificador único de la Whole Slide Image de origen.
patientId	Identificador del paciente al que pertenece la lámina.
fold	Fold de validación cruzada asignado (Val1–Val4).
split	Subconjunto experimental al que pertenece la instancia (entrenamiento o prueba).
gleasonPrimary	Puntaje de Gleason primario asociado a la WSI.
gleasonSecondary	Puntaje de Gleason secundario asociado a la WSI.
isup	Grado ISUP derivado de la combinación de Gleason primario y secundario.
nc, g3, g4, g5, g4c	Indicadores locales de patrones histopatológicos a nivel de parche, utilizados únicamente para trazabilidad y análisis descriptivo.

Es pertinente precisar que, si bien el *manifest* integra descriptores histológicos a nivel de parche, estos atributos se omiten deliberadamente como etiquetas de entrenamiento. El sistema se restringe exclusivamente al uso de etiquetas clínicas globales a nivel de *Whole Slide Image* (WSI), garantizando así el cumplimiento estricto del paradigma de supervisión débil que define al Aprendizaje de Instancias Múltiples y que sustenta la evaluación de los modelos propuestos

7.1.5.3. Justificación Metodológica del Uso de Supervisión Débil

El diseño del pipeline experimental adopta de forma deliberada un esquema de supervisión débil, en el cual las etiquetas clínicas están disponibles únicamente a nivel de lámina completa y no a nivel de parche individual. Esta decisión metodológica responde tanto a consideraciones prácticas como clínicas, ya que en escenarios reales de patología digital las anotaciones exhaustivas a nivel de píxel o parche suelen ser costosas, subjetivas y difícilmente escalables.

El uso de información local únicamente con fines de trazabilidad, y no como señal supervisada directa, evita introducir sesgos artificiales en el aprendizaje y garantiza que el modelo aprenda patrones discriminativos relevantes a partir de la agregación de instancias heterogéneas. De esta manera, el pipeline se alinea con escenarios clínicos realistas, donde la decisión diagnóstica se emite a nivel de lámina, y el sistema debe inferir dicha decisión a partir de múltiples regiones con distintos grados de relevancia histopatológica.

Este enfoque refuerza la validez externa del sistema propuesto y justifica el uso de mecanismos de atención y agregación definidos en etapas posteriores del pipeline.

7.1.6. Preprocesamiento y Control de Calidad de Parches

Antes de la extracción de características, se implementó un protocolo de control de calidad para descartar parches no informativos. Este filtrado consistió en el análisis del contenido tisular mediante la transformación al espacio de color HSV, utilizando específicamente el canal de saturación para discriminar regiones con

presencia significativa de tejido.

Aquellos parches que no alcanzaron un umbral predefinido de proporción tisular fueron excluidos para mitigar el ruido experimental y optimizar la carga computacional. Este procedimiento se realizó de forma agnóstica a las etiquetas clínicas, garantizando la integridad del experimento y evitando el sesgo por fuga de información (data leakage).

7.1.7. Extracción de Características Profundas

Con el objetivo de desacoplar la representación visual del entrenamiento de los modelos de Aprendizaje por Instancias Múltiples (MIL), se implementó una fase de extracción de características profundas.

Cada parche válido fue procesado mediante una arquitectura ResNet-50 preentrenada en ImageNet, empleada como extractor de características (feature extractor) tras la remoción de su capa de clasificación final (top layer). De este modo, cada parche se transformó en un vector de características (embedding) de dimensión fija, diseñado para codificar patrones morfológicos complejos de la arquitectura tisular y celular.

Estos descriptores se almacenaron de forma persistente, optimizando el flujo de trabajo al permitir su reutilización en diversas configuraciones experimentales sin incurrir en el costo computacional de reprocesar las imágenes originales.

7.1.8. Construcción Formal de Bolsas MIL

Utilizando el registro centralizado en `dataset_manifest.csv`, se estructuraron los datos bajo el paradigma de Aprendizaje por Instancias Múltiples (MIL). En este esquema, cada bolsa (*bag*) se define como una imagen de lámina completa (WSI), mientras que cada parche histopatológico constituye una instancia dentro de dicha bolsa.

La etiqueta de supervisión de la bolsa se asignó exclusivamente a partir del grado ISUP clínico de la WSI original. Por su parte, las instancias mantuvieron sus atributos locales (coordenadas y descriptores) con fines estrictamente descriptivos y de trazabilidad, sin participar en la supervisión directa durante el entrenamiento.

Se implementó un protocolo de validación para garantizar la integridad estructural de las bolsas, asegurando que cada una contuviera exclusivamente instancias provenientes de su WSI correspondiente y que no existiera contaminación cruzada entre particiones experimentales. Esta configuración preserva la cardinalidad variable de las bolsas, característica estructural inherente al problema MIL en histopatología digital, donde la cantidad de regiones informativas puede variar sustancialmente entre láminas. Mantener esta variabilidad resulta esencial para evaluar de manera realista la capacidad de los mecanismos de agregación para adaptarse a distribuciones espaciales heterogéneas de patrones tisulares.

7.1.9. Estrategia experimental para el modelado MIL binario

Sobre los bags MIL construidos, se definió un problema de clasificación binaria a nivel de WSI, cuyo objetivo consiste en discriminar entre casos clínicamente relevantes y no relevantes a partir del grado ISUP asociado a cada lámina.

Se implementaron cuatro arquitecturas MIL ampliamente utilizadas en la literatura:

- Mean Pooling MIL

- Max Pooling MIL
- Attention-based MIL (ABMIL)
- Gated Attention MIL (SmABMIL)

Las representaciones de cada parche fueron obtenidas previamente mediante una arquitectura ResNet-50 preentrenada en ImageNet, utilizada exclusivamente como extractor de características. La capa de clasificación final fue removida y todos los pesos se mantuvieron congelados durante el entrenamiento de los modelos MIL, con el fin de reducir el riesgo de sobreajuste dado el número limitado de WSIs disponibles.

El entrenamiento se realizó durante 10 épocas por modelo y por fold, empleando validación cruzada estricta a nivel de WSI, sin solapamiento entre conjuntos de entrenamiento y prueba. Las métricas clínicas se calcularon exclusivamente a nivel de bolsa (WSI-level), en coherencia con la formulación del problema bajo el paradigma MIL.

7.1.10. Resumen de la Metodología Experimental

La metodología expuesta consolida un marco experimental caracterizado por su rigor, reproducibilidad y fundamentación clínica para la evaluación de modelos MIL en histopatología prostática. La separación explícita entre la preparación de datos, el control de integridad, la extracción de descriptores y la fase de entrenamiento garantiza la ortogonalidad de los procesos. Este diseño asegura que el desempeño observado sea atribuible exclusivamente al comportamiento intrínseco de las arquitecturas evaluadas, eliminando posibles sesgos o artefactos derivados del preprocesamiento.

La adopción del `dataset_manifest.csv` como **registro maestro de datos**, en conjunto con el paradigma de supervisión débil y un particionamiento estrictamente independiente a nivel de paciente (*patient-level partitioning*), define un entorno experimental homogéneo, auditable y libre de fuga de información (*data leakage*). Bajo estas condiciones controladas, las métricas obtenidas reflejan de manera fiel la capacidad de generalización de los modelos ante la variabilidad morfológica no observada durante el entrenamiento, un factor crítico en el despliegue de herramientas de apoyo al diagnóstico.

Este andamiaje metodológico no solo sustenta la validez de los experimentos presentados en el capítulo de Resultados, sino que también delimita el marco interpretativo para la Discusión. En particular, proporciona la base necesaria para evaluar la robustez y la viabilidad clínica del enfoque propuesto, garantizando que las conclusiones extraídas posean la solidez técnica requerida en el ámbito de la patología digital.

7.2. Construcción y validación del dataset experimental

A partir del conjunto original de imágenes histopatológicas SICAPv2, se construyó un archivo *dataset_manifest.csv* que actúa como única fuente de verdad para todos los experimentos posteriores. En este manifest, cada fila corresponde a un parche histopatológico y queda asociado de forma explícita a su Whole Slide Image (WSI), fold de validación cruzada, subconjunto (entrenamiento o prueba), rutas a imagen y máscara, así como a los labels clínicos globales derivados a nivel WSI.

El manifest final contiene un total de **39,836** parches, organizados en **496** WSIs, distribuidos en cuatro folds de validación cruzada (Val1–Val4). Cada WSI posee un único conjunto de labels clínicos (Gleason primario, Gleason secundario e ISUP), los cuales se derivan una sola vez a nivel WSI y posteriormente se

propagan a todos los parches asociados, garantizando coherencia clínica y evitando inconsistencias a nivel de instancia.

Durante la construcción del manifest se realizaron validaciones automáticas críticas orientadas a garantizar la integridad estructural, la coherencia clínica y la trazabilidad completa de los datos experimentales. En particular, se verificó que:

- Ausencia de valores nulos en columnas clave.
- Existencia y trazabilidad de las máscaras asociadas a los parches.
- Consistencia estricta de los labels ISUP a nivel WSI (un único valor ISUP por WSI).

Todas las validaciones fueron superadas satisfactoriamente, por lo que el archivo *dataset_manifest.csv* se considera estable, coherente y clínicamente consistente, y se utiliza como base única para la construcción de bags MIL, el análisis estructural y el entrenamiento de los modelos.

7.2.1. Preparación de Bags MIL y análisis estructural

A partir del *dataset_manifest.csv*, se construyó una representación basada en Aprendizaje de Instancias Múltiples (Multiple Instance Learning, MIL), donde **cada bag corresponde a una WSI** y **cada instancia corresponde a un parche histopatológico**. El label del bag se definió como el grado ISUP asociado a la WSI, mientras que las instancias conservan información local relacionada con la presencia de patrones histológicos (NC, G3, G4, G5 y G4C), utilizada únicamente con fines de trazabilidad.

El conjunto final contiene un total de **496 bags MIL**, distribuidos de forma consistente entre folds y subconjuntos de entrenamiento y prueba. Se verificó que:

- Cada bag contiene exclusivamente instancias pertenecientes a una única WSI.
- No existe solapamiento entre bags de entrenamiento y prueba.
- Todos los bags presentan un único label ISUP válido.
- No existen bags vacíos ni instancias huérfanas.

7.2.2. Distribución de instancias por bag

El número de instancias por bag presenta una variabilidad considerable, reflejando la heterogeneidad espacial inherente a las imágenes histopatológicas. En particular, se observa:

- Número mínimo de instancias por bag: 7
- Número máximo de instancias por bag: 347
- Promedio de instancias por bag: 80.31

La Figura 7.2 ilustra la distribución del número de instancias por WSI, evidenciando una marcada asimetría hacia la derecha. La mayoría de las WSIs presentan un número bajo o intermedio de parches, mientras que un subconjunto reducido concentra una cantidad significativamente mayor de instancias. Este comportamiento refleja la heterogeneidad espacial inherente a la histopatología digital, donde la extensión del tejido

tumoral, la variabilidad morfológica y el tamaño efectivo de la región anotada pueden diferir considerablemente entre pacientes. Desde el punto de vista metodológico, esta variabilidad justifica plenamente el uso del paradigma MIL, el cual permite manejar bags de tamaño variable sin imponer restricciones artificiales sobre el número de instancias, preservando así la información clínica y morfológica relevante contenida en cada WSI.

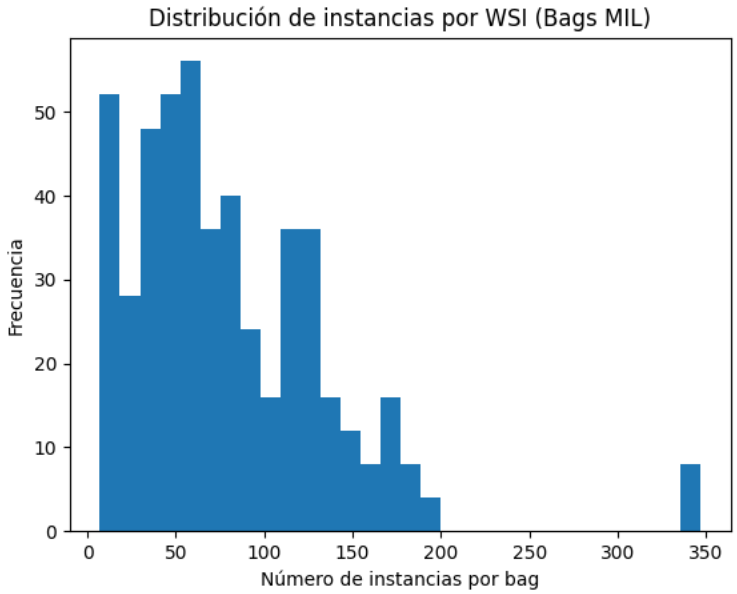

Figura 7.2: Distribución del número de instancias por WSI. Se observa una marcada asimetría hacia la derecha.

7.2.3. Dinámica de entrenamiento y convergencia

La evolución de la función de pérdida durante el entrenamiento se analizó tanto de forma agregada (promedio sobre folds) como individualmente por fold y modelo.

7.2.3.1. Curvas de pérdida promedio

La Figura 7.3 muestra la evolución promedio de la pérdida (Binary Cross-Entropy) para los cuatro modelos MIL a lo largo de las épocas de entrenamiento. Se observa que:

- **MeanMIL** presenta una convergencia estable pero relativamente lenta.
- **MaxMIL** exhibe inestabilidad en las primeras épocas, seguida de una reducción progresiva de la pérdida.
- **ABMIL** y **SmABMIL** muestran una convergencia más pronunciada y consistente, con reducciones sustanciales de la pérdida desde las primeras épocas.

Estas diferencias en las curvas de pérdida sugieren que los modelos basados en mecanismos de atención son capaces de identificar y ponderar de manera más eficiente las instancias relevantes dentro de cada bag

desde etapas tempranas del entrenamiento. En consecuencia, el proceso de optimización resulta más estable y consistente, reduciendo la variabilidad inter-época observada en los enfoques basados exclusivamente en operaciones de pooling.

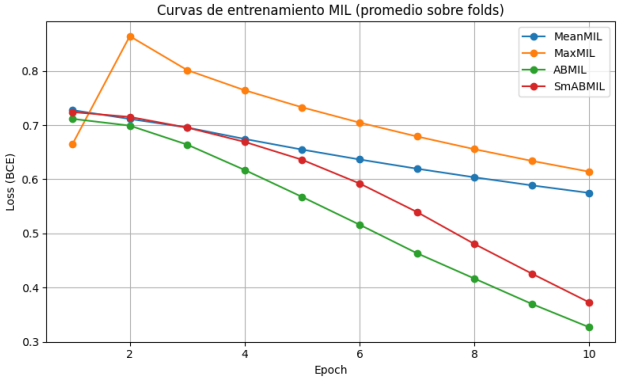

Figura 7.3: Curvas de pérdida promedio (Binary Cross-Entropy) por modelo MIL. Se observa una convergencia más rápida y estable en los modelos basados en atención (ABMIL y SmABMIL).

7.2.3.2. Curvas de pérdida por fold

Con el fin de analizar la estabilidad del proceso de entrenamiento y la robustez de cada arquitectura frente a distintas particiones del conjunto de datos, se examinaron las curvas de pérdida (*Binary Cross-Entropy*) de manera individual para cada fold de validación cruzada y cada modelo MIL considerado.

Este análisis resulta particularmente relevante en el contexto de histopatología digital basada en *Whole Slide Images*, donde la heterogeneidad inter-WSI y la variabilidad en el número de instancias por bag pueden inducir comportamientos de entrenamiento distintos dependiendo de la composición específica de cada fold.

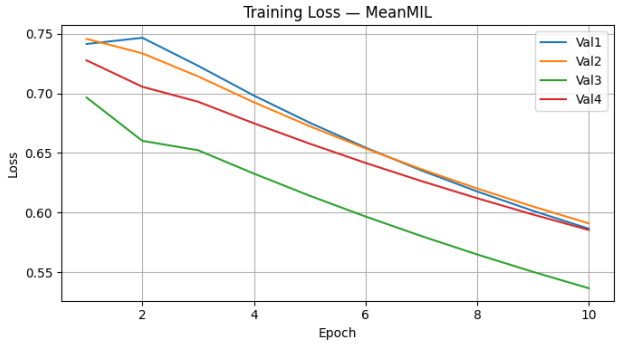

Figura 7.4: Curvas de pérdida por fold para el modelo Mean Pooling MIL. Cada curva representa la evolución de la pérdida durante el entrenamiento en un fold distinto de validación cruzada.

En el caso de **MeanMIL** (Figura 7.4), se observa una convergencia relativamente estable en los cuatro

folds, aunque con diferencias moderadas en la pendiente inicial y en el valor final de la pérdida. Este comportamiento refleja la naturaleza del promedio como operador de agregación, el cual incorpora información de todas las instancias del bag de forma uniforme, pero puede verse afectado por la presencia de instancias no informativas o ruidosas, particularmente en WSIs con alta heterogeneidad histológica.

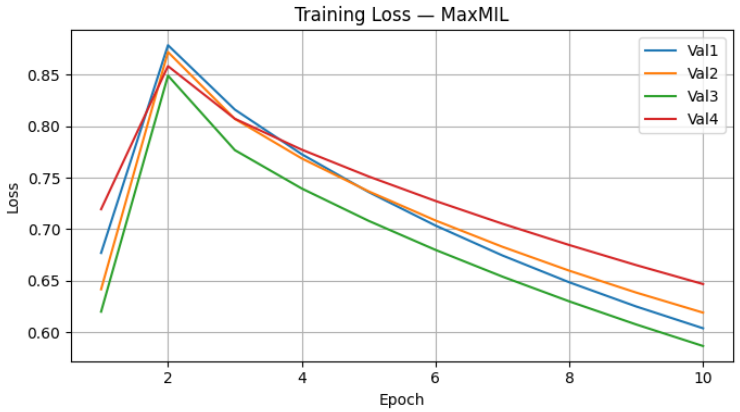

Figura 7.5: Curvas de pérdida por fold para el modelo Max Pooling MIL. Se evidencia una mayor variabilidad entre folds durante las primeras épocas de entrenamiento.

Para el modelo **MaxMIL** (Figura 7.5), las curvas de pérdida muestran una mayor sensibilidad a la partición de los datos, especialmente durante las primeras épocas. Esta inestabilidad es consistente con el mecanismo de max pooling, que basa la predicción del bag en una única instancia dominante, lo cual puede amplificar el impacto de instancias atípicas o poco representativas en determinados folds.

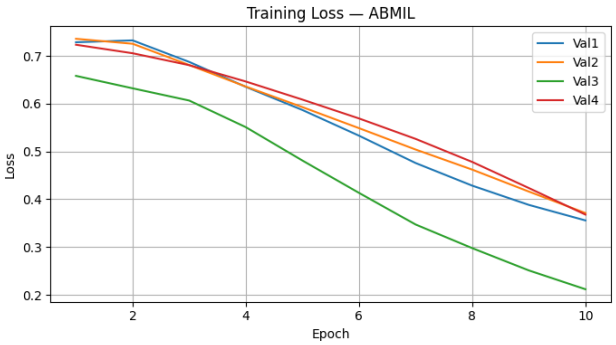

Figura 7.6: Curvas de pérdida por fold para el modelo Attention-based MIL (ABMIL). Se observa una convergencia rápida y consistente entre folds.

El modelo **ABMIL** (Figura 7.6) presenta un patrón de convergencia más uniforme entre folds, con reducciones pronunciadas de la pérdida desde las primeras épocas y valores finales similares. Este comportamiento sugiere que el mecanismo de atención aprende de manera efectiva a ponderar las instancias más relevantes dentro de cada bag, mitigando el efecto del ruido y favoreciendo una optimización más estable independien-

temente de la partición de los datos.

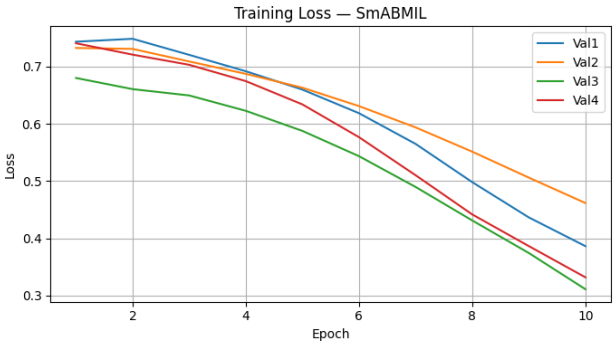

Figura 7.7: Curvas de pérdida por fold para el modelo Gated Attention MIL (SmABMIL). La convergencia es consistente y ligeramente más estable que en ABMIL.

Finalmente, el modelo **SmABMIL** (Figura 7.7) exhibe el comportamiento más estable entre folds, con trayectorias de pérdida altamente consistentes y una convergencia suave. La incorporación del mecanismo de compuertas (*gated attention*) permite modelar interacciones más complejas entre instancias, lo cual se traduce en un proceso de entrenamiento menos sensible a la variabilidad estructural entre WSIs y folds.

Este análisis confirma que los modelos basados en mecanismos de atención no solo convergen más rápidamente, sino que también presentan una mayor robustez frente a las diferentes particiones de validación cruzada, un aspecto clave en escenarios clínicos donde la variabilidad interpaciente y la heterogeneidad tisular son inherentes al problema.

7.2.4. Resultados clínicos cuantitativos por modelo y fold

La evaluación del desempeño de los modelos de *Multiple Instance Learning* se realizó de manera estricta a nivel de *Whole Slide Image* (WSI), en coherencia con la formulación del problema, el paradigma MIL adoptado y las métricas definidas en la Sección correspondiente. En ningún caso se evaluaron predicciones a nivel de parche.

Para cada fold de validación cruzada (`Val1-Val4`), se entrenaron y evaluaron de forma independiente los cuatro modelos considerados: **MeanMIL**, **MaxMIL**, **ABMIL** y **SmABMIL**. Las métricas reportadas incluyen *accuracy*, sensibilidad, especificidad, precisión, F1-score y AUC-ROC, todas calculadas sobre el conjunto de prueba de cada fold.

Fold Val1

Cuadro 7.2: Resultados clínicos a nivel WSI para el fold Val1

Modelo	Acc	Sens	Spec	Prec	F1	AUC
MeanMIL	0.690	0.625	1.000	1.000	0.769	0.850
MaxMIL	0.172	0.000	1.000	0.000	0.000	0.942
ABMIL	0.724	0.792	0.400	0.864	0.826	0.708
SmABMIL	0.897	0.917	0.800	0.957	0.936	0.975

Fold Val2

Cuadro 7.3: Resultados clínicos a nivel WSI para el fold Val2

Modelo	Acc	Sens	Spec	Prec	F1	AUC
MeanMIL	0.815	0.850	0.714	0.895	0.872	0.914
MaxMIL	0.296	0.050	1.000	1.000	0.095	0.900
ABMIL	0.815	0.900	0.571	0.857	0.878	0.836
SmABMIL	0.889	0.950	0.714	0.905	0.927	0.886

Fold Val3

Cuadro 7.4: Resultados clínicos a nivel WSI para el fold Val3

Modelo	Acc	Sens	Spec	Prec	F1	AUC
MeanMIL	0.600	0.786	0.438	0.647	0.710	0.661
MaxMIL	0.600	0.143	1.000	1.000	0.250	0.612
ABMIL	0.633	0.786	0.500	0.688	0.733	0.661
SmABMIL	0.633	0.786	0.500	0.688	0.733	0.746

Fold Val4

Cuadro 7.5: Resultados clínicos a nivel WSI para el fold Val4

Modelo	Acc	Sens	Spec	Prec	F1	AUC
MeanMIL	0.711	1.000	0.154	0.694	0.820	0.975
MaxMIL	0.526	0.280	1.000	1.000	0.438	0.938
ABMIL	0.842	0.920	0.692	0.885	0.902	0.945
SmABMIL	0.711	0.960	0.231	0.706	0.813	0.818

Resumen estadístico global

Los resultados clínicos cuantitativos obtenidos para cada fold de validación cruzada permiten evaluar no solo el desempeño promedio de los modelos, sino también su estabilidad frente a variaciones en la partición de los datos. Dado el número limitado de WSIs y la heterogeneidad intrínseca del tejido prostático, esta evaluación por fold resulta particularmente relevante en un contexto clínico.

En la siguiente tabla se reportan métricas ampliamente utilizadas en clasificación médica, donde la sensibilidad refleja la capacidad del modelo para identificar correctamente casos clínicamente relevantes, mientras que la especificidad cuantifica la correcta identificación de casos no relevantes. Métricas como F1-score y AUC-ROC permiten evaluar el equilibrio global entre detección y precisión, proporcionando una visión integral del comportamiento diagnóstico de cada arquitectura.

Cuadro 7.6: Resumen estadístico (media ± desviación estándar) de métricas clínicas a nivel WSI

Modelo	**Acc**	**Sens**	**Spec**	**Prec**	**F1**	**AUC**
MeanMIL	0.704±0.088	0.815±0.155	0.576±0.363	0.785±0.201	0.777±0.096	0.850±0.136
MaxMIL	0.399±0.199	0.118±0.123	1.000±0.000	0.750±0.500	0.196±0.191	0.848±0.159
ABMIL	0.754±0.095	0.849±0.071	0.541±0.123	0.788±0.139	0.814±0.102	0.787±0.128
SmABMIL	0.782±0.131	0.903±0.080	0.561±0.254	0.787±0.176	0.836±0.126	0.856±0.098

El resumen estadístico global pone de manifiesto diferencias claras entre las arquitecturas evaluadas. En particular, los modelos basados en atención muestran, en promedio, valores superiores de sensibilidad y F1-score, así como una menor variabilidad inter-fold, lo cual sugiere un comportamiento más robusto frente a la heterogeneidad de las WSIs.

Por el contrario, el modelo MaxMIL presenta una alta especificidad acompañada de una sensibilidad consistentemente baja, indicando una tendencia a clasificar correctamente los casos negativos a costa de omitir una proporción significativa de casos positivos, un comportamiento clínicamente desfavorable en escenarios de detección de patología.

7.2.5. Matrices de confusión promedio por modelo

Con el objetivo de complementar los resultados cuantitativos reportados en la sección anterior, se presentan las matrices de confusión promedio obtenidas para cada modelo de *Multiple Instance Learning*. Estas matrices se construyeron agregando los resultados de los cuatro folds de validación cruzada y se reportan exclusivamente a nivel de *Whole Slide Image* (WSI).

Las matrices de confusión promedio proporcionan una representación directa y clínicamente interpretable del comportamiento de cada modelo, al descomponer explícitamente las predicciones correctas e incorrectas a nivel de WSI. En un contexto diagnóstico, este tipo de análisis resulta fundamental, ya que permite evaluar no solo el desempeño global, sino también la naturaleza de los errores cometidos.

En particular, la relación entre falsos negativos y verdaderos positivos adquiere especial relevancia clínica, dado que un falso negativo implica la omisión de un caso patológico potencialmente significativo. Las Figuras correspondientes permiten observar patrones diferenciados entre modelos, evidenciando cómo las distintas

estrategias de agregación de instancias influyen directamente en la distribución de errores y aciertos.

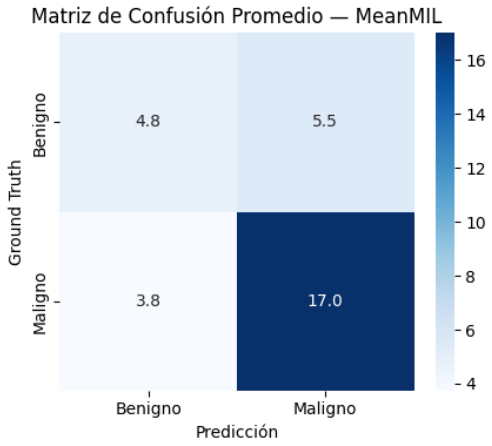

Figura 7.8: Matriz de confusión promedio a nivel WSI para el modelo MeanMIL.

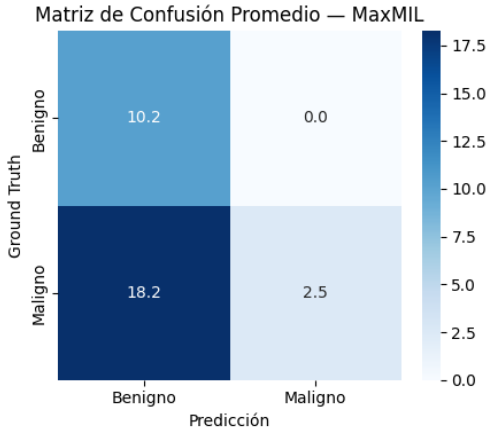

Figura 7.9: Matriz de confusión promedio a nivel WSI para el modelo MaxMIL.

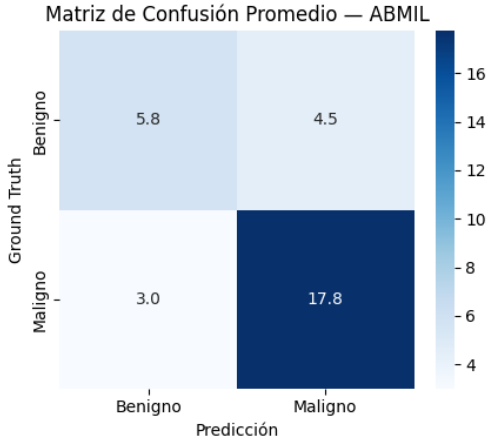

Figura 7.10: Matriz de confusión promedio a nivel WSI para el modelo ABMIL.

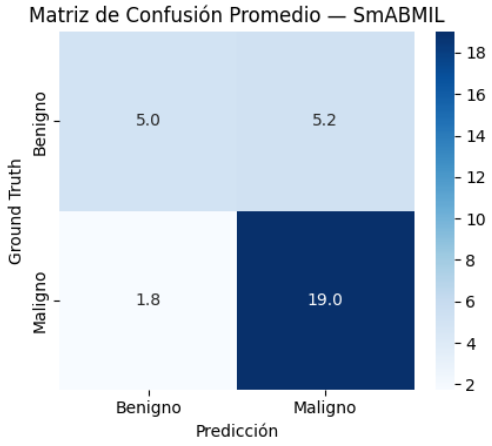

Figura 7.11: Matriz de confusión promedio a nivel WSI para el modelo SmABMIL.

Cuadro 7.7: Resumen de matrices de confusión (media ± desviación estándar) por modelo

Modelo	**TN**	**FP**	**FN**	**TP**
MeanMIL	$4{,}75 \pm 2{,}06$	$5{,}50 \pm 5{,}32$	$3{,}75 \pm 3{,}77$	$17{,}00 \pm 5{,}89$
MaxMIL	$10{,}25 \pm 5{,}12$	$0{,}00 \pm 0{,}00$	$18{,}25 \pm 4{,}92$	$2{,}50 \pm 3{,}11$
ABMIL	$5{,}75 \pm 3{,}30$	$4{,}50 \pm 2{,}38$	$3{,}00 \pm 1{,}41$	$17{,}75 \pm 4{,}99$
SmABMIL	$5{,}00 \pm 2{,}16$	$5{,}25 \pm 4{,}43$	$1{,}75 \pm 0{,}96$	$19{,}00 \pm 5{,}72$

Los resultados obtenidos evidencian diferencias claras entre las arquitecturas MIL evaluadas, tanto en términos de estabilidad de entrenamiento como de desempeño clínico a nivel WSI. En el siguiente capítulo, estos hallazgos se analizan en profundidad, relacionándolos con las decisiones metodológicas adoptadas, las características estructurales del dataset y las implicaciones clínicas del uso de mecanismos de atención.

Capítulo 8

Análisis de Resultados y Mapas de Atención

Este capítulo integra de manera unificada los resultados derivados de la etapa de procesamiento de datos, el análisis estructural del conjunto histopatológico y la evaluación cuantitativa y cualitativa de las arquitecturas basadas en *Multiple Instance Learning* (MIL). La exposición sigue una progresión lógica que parte de la consolidación del dataset, continúa con la caracterización estadística de su estructura y culmina con la interpretación metodológica y clínica del comportamiento de los modelos propuestos.

8.1. Resultados de la Etapa de Procesamiento y Consolidación del Dataset

La construcción del archivo maestro `dataset_manifest.csv` permitió estructurar de forma consistente la totalidad del conjunto SICAPv2 bajo un esquema reproducible y clínicamente coherente. El dataset consolidado contiene un total de 39 836 parches histopatológicos asociados a 124 *Whole Slide Images* (WSIs) únicas. Cada instancia quedó vinculada a su identificador de WSI, partición oficial (Val1–Val4), subconjunto (train/test), metadatos clínicos y grado ISUP derivado a nivel global.

Las validaciones programáticas confirmaron la integridad estructural del conjunto: no se detectaron valores nulos, inconsistencias imagen–máscara ni variabilidad intra-WSI del grado ISUP. Asimismo, se verificó ausencia total de solapamiento entre entrenamiento y prueba en cada partición oficial, garantizando estrictamente la no existencia de *data leakage* a nivel de paciente.

Desde el punto de vista clínico, la distribución del grado ISUP evidenció una mayor proporción de casos clínicamente significativos. Tras la binarización definida como ISUP ≥ 2 (malignidad clínicamente relevante) frente a ISUP $= 1$ (benigno), la composición final a nivel WSI fue de 332 casos malignos y 164 benignos. Esta distribución, equivalente a aproximadamente 66.9 % de casos positivos, introduce un desbalance moderado que resulta representativo de cohortes clínicas orientadas a diagnóstico oncológico.

8.2. Construcción y Análisis Estructural de las Bolsas MIL

A partir del manifest consolidado se construyó la representación MIL formal, donde cada WSI constituye un *bag* y cada parche una instancia. Dado que el protocolo experimental emplea cuatro particiones oficiales de validación cruzada, las 124 WSIs originales generan un total de 496 bolsas MIL (124 WSIs $\times$ 4 folds), manteniendo separación estricta entre entrenamiento y prueba en cada partición.

El análisis estadístico del tamaño de las bolsas reveló una heterogeneidad estructural considerable. El número mínimo de instancias por bolsa fue 7, mientras que el máximo alcanzó 347 parches por WSI. La media fue de 80.31 instancias, con una desviación estándar de 57.08. Esta dispersión elevada respecto a la

media confirma una variabilidad sustancial en la extensión espacial del tejido representado por cada biopsia digitalizada.

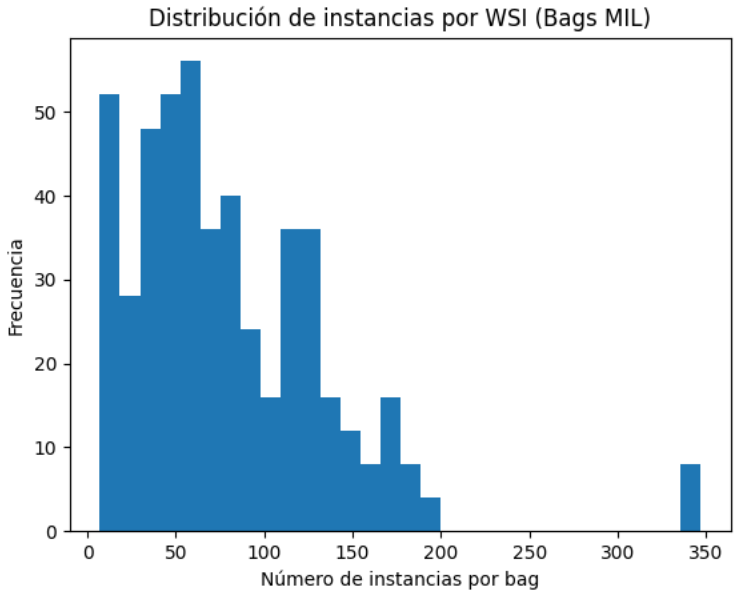

Figura 8.1: Distribución del número de instancias por bolsa MIL. Se evidencia alta heterogeneidad estructural entre WSIs.

Desde una perspectiva metodológica, esta heterogeneidad justifica el uso de enfoques MIL con mecanismos de agregación adaptativa. La amplia variabilidad en el tamaño de los bags implica que estrategias de pooling uniforme pueden diluir señales tumorales focales o amplificar instancias no representativas. Por tanto, la estructura estadística del dataset no solo describe su composición, sino que fundamenta la elección arquitectónica adoptada.

8.3. Evaluación Cuantitativa y Comparación de Arquitecturas MIL

La comparación entre las arquitecturas evaluadas (MeanMIL, MaxMIL, ABMIL y SmABMIL) evidenció diferencias consistentes asociadas a sus mecanismos de agregación. Los métodos no adaptativos mostraron limitaciones estructurales frente a la heterogeneidad espacial del tejido prostático. En particular, MeanMIL tendió a promediar uniformemente regiones informativas con tejido benigno dominante, reduciendo la sensibilidad frente a lesiones focales. MaxMIL, por su parte, dependió exclusivamente de la instancia con activación máxima, generando un comportamiento inestable y clínicamente riesgoso al incrementar la probabilidad de falsos negativos en presencia de ruido o artefactos locales.

En contraste, las arquitecturas basadas en atención (ABMIL y especialmente SmABMIL) demostraron mayor robustez y estabilidad inter-fold. El mecanismo de atención permitió asignar pesos diferenciados a subconjuntos informativos dentro del bag, adaptando la representación global a la distribución real de patrones histológicos. La variante SmABMIL, mediante su mecanismo de atención compuerta, actuó además como regularizador implícito, evitando concentraciones excesivas en una sola instancia y favoreciendo representaciones más balanceadas.

Este comportamiento se tradujo en un compromiso clínicamente más adecuado entre sensibilidad y especificidad, priorizando la detección de casos malignos sin degradar sustancialmente el control de falsos positivos. Dado el predominio de casos clínicamente significativos en la cohorte, la optimización de la sensibilidad adquiere especial relevancia en escenarios de cribado asistido.

8.4. Interpretabilidad y Correlato Histopatológico

Más allá de las métricas globales, el análisis de interpretabilidad constituye un componente central para la integración clínica del sistema. La distribución de probabilidades predichas por SmABMIL mostró una marcada polarización hacia valores cercanos a 0 y 1, con baja densidad en el rango intermedio de ambigüedad, lo que sugiere alta confianza predictiva y estabilidad decisional.

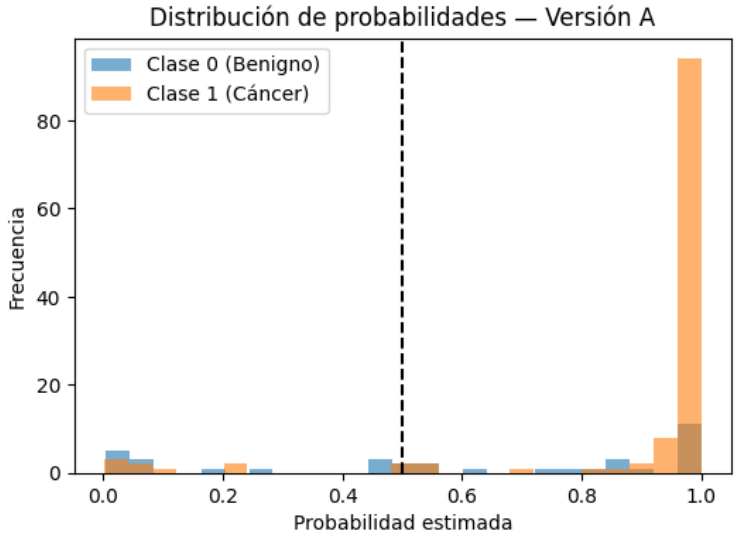

Figura 8.2: Distribución de probabilidades predichas por SmABMIL. Se observa polarización hacia valores extremos, indicativa de alta confianza predictiva.

Adicionalmente, la inspección de los parches con mayor peso de atención (*Top-k*) evidenció coherencia morfológica con criterios diagnósticos reconocidos del cáncer de próstata. Las regiones resaltadas correspondieron predominantemente a áreas con arquitectura glandular alterada, fusión de glándulas y pérdida de organización estructural, mientras que el tejido estromal conservado recibió pesos marginales.

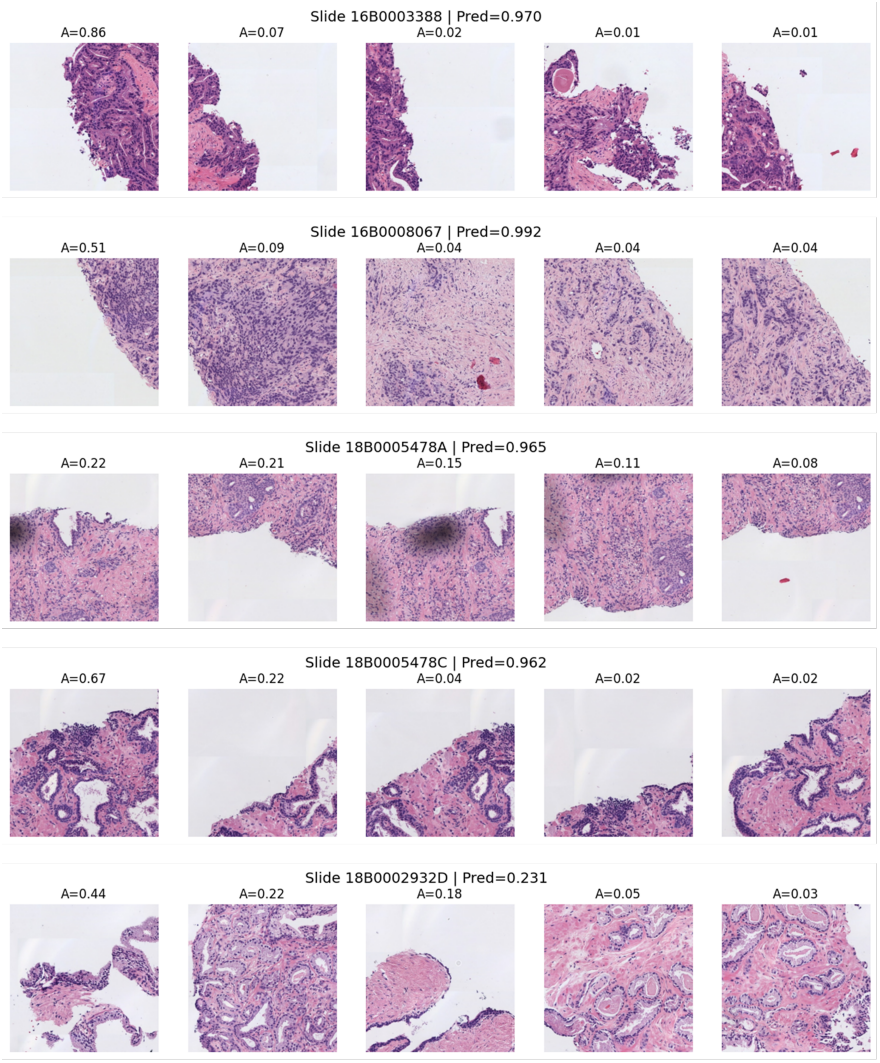

Figura 8.3: Parches con mayor peso de atención (*Top-k*). El modelo focaliza regiones morfológicamente compatibles con malignidad bajo supervisión débil.

Este correlato histopatológico emergió bajo un esquema de supervisión débil, utilizando exclusivamente etiquetas a nivel WSI. La capacidad del modelo para identificar regiones clínicamente relevantes sin anotaciones locales manuales refuerza su potencial como herramienta de apoyo diagnóstico y reduce el carácter opaco asociado tradicionalmente a los sistemas de aprendizaje profundo.

8.5. Consideraciones Metodológicas y Proyección

A pesar de la solidez estructural y del desempeño observado, el estudio presenta limitaciones inherentes al manejo de imágenes gigapíxel y al tamaño variable de las bolsas, lo cual impone restricciones computacionales y de memoria en GPU. Asimismo, la formulación binaria del problema simplifica la complejidad clínica

real del sistema ISUP multiclase.

No obstante, los resultados obtenidos demuestran de forma consistente que la combinación de una estructuración rigurosa del dataset, una representación MIL adecuada a la heterogeneidad espacial y un mecanismo de atención compuerta constituye una base metodológica sólida para el desarrollo de sistemas de apoyo diagnóstico en patología prostática digital.

En síntesis, la etapa de procesamiento no solo garantizó integridad y coherencia estructural del conjunto de datos, sino que estableció las condiciones estadísticas y clínicas que justifican la superioridad de los modelos MIL basados en atención, consolidando una transición metodológicamente fundamentada hacia su aplicación en escenarios diagnósticos reales.

Capítulo 9

Conclusiones y Trabajo Futuro

El presente proyecto abordó el problema de la clasificación de cáncer de próstata a partir de imágenes histopatológicas digitalizadas mediante un enfoque basado en Aprendizaje de Instancias Múltiples (MIL). A lo largo de su desarrollo se diseñó, implementó y evaluó un pipeline experimental integral y reproducible, que abarca desde la preparación rigurosa de los datos hasta el análisis cuantitativo del desempeño y la interpretabilidad visual de los modelos entrenados. En función de los objetivos planteados y de los resultados obtenidos, se derivan las siguientes conclusiones principales.

9.1. Conclusiones

- **Viabilidad del Aprendizaje con Supervisión Débil en Histopatología Digital:** Los resultados confirman que es factible entrenar modelos con capacidad discriminativa clínicamente relevante utilizando exclusivamente etiquetas a nivel de *Whole Slide Image* (WSI), sin requerir anotaciones exhaustivas a nivel de parche. El paradigma MIL permitió capturar patrones tisulares asociados a malignidad prostática de manera efectiva, validando su idoneidad para escenarios clínicos reales donde la anotación detallada resulta costosa, subjetiva o impracticable.

- **Superioridad de los Modelos MIL Basados en Atención:** La comparación sistemática entre estrategias de agregación evidenció que los modelos basados en mecanismos de atención (ABMIL y SmABMIL) superan de forma consistente a los enfoques de pooling clásico (MeanMIL y MaxMIL) en métricas clínicas clave, incluyendo sensibilidad, PR AUC y estabilidad inter-fold. En particular, SmABMIL mostró un equilibrio favorable entre capacidad predictiva y robustez, lo que sugiere una mejor generalización frente a la heterogeneidad morfológica inherente a las biopsias prostáticas.

- **Impacto Crítico del Preprocesamiento y la Curaduría de Datos:** Las decisiones adoptadas en las etapas iniciales del pipeline, tales como el filtrado basado en el espacio de color HSV y la selección de parches según cobertura tisular, tuvieron un impacto directo sobre la estabilidad del entrenamiento y la calidad de las representaciones aprendidas. Estos procesos resultaron fundamentales para reducir ruido estructural y garantizar que el modelo se enfocara en regiones con relevancia histopatológica, condicionando de manera significativa el desempeño final del sistema.

- **Rigor Metodológico y Evaluación Realista del Desempeño:** La implementación de un esquema de validación cruzada estricta mediante *GroupKFold*, asegurando la separación completa a nivel de paciente entre los conjuntos de entrenamiento y prueba, permitió evitar fugas de información y obtener estimaciones realistas de la capacidad de generalización del modelo. Este rigor metodológico constituye un requisito indispensable para la evaluación confiable de sistemas de apoyo clínico basados en aprendizaje automático.

- **Interpretabilidad y Coherencia con Criterios Histopatológicos:** La incorporación de mecanismos de atención proporcionó una capa de interpretabilidad esencial para el contexto clínico. La visualización de los pesos de atención, tanto a nivel de parches individuales como mediante mapas

espaciales proyectados sobre la WSI, evidenció una correspondencia coherente con regiones tisulares relevantes desde el punto de vista histopatológico. Esta capacidad de explicación fortalece la confianza en el sistema y lo posiciona como una herramienta de apoyo clínico interpretable, más allá de un clasificador de tipo caja negra.

9.2. Trabajo Futuro

Si bien los resultados obtenidos son prometedores, existen diversas líneas de investigación que permitirían ampliar el alcance, la robustez y el impacto clínico del sistema propuesto:

- **Aprendizaje Auto-Supervisado y Representaciones Especializadas:** Explorar extractores de características entrenados mediante aprendizaje auto-supervisado (SSL), como DINO o MoCo, así como arquitecturas basadas en *Vision Transformers* (ViT), con el objetivo de obtener representaciones más específicas para la morfología histopatológica que aquellas preentrenadas en conjuntos genéricos como ImageNet.
- **Análisis Multi-Resolución:** Incorporar esquemas MIL multi-escala que procesen parches a diferentes aumentos (10×, 20×, 40×), replicando de manera más fiel el flujo de trabajo diagnóstico del patólogo e integrando información arquitectónica global con detalles citonucleares locales.
- **Integración Multimodal y Validación Externa:** Extender el sistema mediante la fusión de imágenes histopatológicas con variables clínicas relevantes, como PSA, edad o resultados de resonancia magnética, así como validar el modelo en *datasets* externos de distintas instituciones, fortaleciendo su robustez frente a variaciones de protocolo, tinción y población.
- **Aplicaciones Clínicas Interactivas:** Desarrollar integraciones con visores de WSI de código abierto, como QuPath, que permitan utilizar el modelo como herramienta de apoyo a la decisión clínica en tiempo real, resaltando regiones sospechosas directamente sobre la lámina digital y facilitando su adopción en entornos clínicos reales.

Capítulo 10

Anexos

- Cronograma de actividades
- Cuaderno de Implementación del Modelo Attention-MIL y Flujo de Ingeniería de Datos.
- Repositorio de Github del Proyecto de Investigación