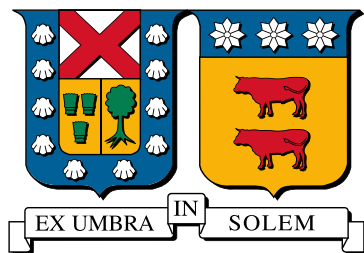


UNIVERSIDAD TÉCNICA FEDERICO SANTA
MARÍA

DEPARTAMENTO DE ELECTRÓNICA

VALPARAÍSO - CHILE



“SISTEMA MULTIAGENTE DE
INTELIGENCIA ARTIFICIAL PARA
ONCOLOGÍA DE PRECISIÓN”

SEBASTIÁN ANDRÉS DONOSO ARQUERO

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL
ELECTRÓNICO.

PROFESOR GUIA: DR. WERNER CREIXELL FUENTES
PROFESOR CORREFERENTE: DR. ALEJANDRO WEINSTEIN
OPPENHEIMER

MARZO 2026



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): Memoria o trabajo de título Tesis de Postgrado

Título del trabajo: Sistema Multiagente De Inteligencia Artificial Para Oncología de Precisión

Nombre del candidato(a): Sebastián Andrés Donoso Arquero

Carrera / Grado: Ingeniería Civil Electrónica

Campus: Casa Central Departamento: Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Werner Geixell, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (marcar una opción):

6 meses 12 meses 2 años 3 años 5 años 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 14/05/2026 Firma: Werner Geixell

Estudiante o Candidato(a):

Fecha: 14-05-2026 Firma: Sebastián Andrés Donoso Arquero

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Sistema Multiagente de Inteligencia Artificial para

Oncología de Precisión

Sebastián Andrés Donoso Arquero

Memoria para optar al título de Ingeniero Civil Electrónico, mención Estructuras y
Sistemas Computacionales, submención Informática.

Universidad Técnica Federico Santa María

Profesor Guía: Dr. Werner Creixell Fuentes

Profesor Correferente: Dr. Alejandro Weinstein

MARZO 2026

Resumen

El diagnóstico de cáncer de mama mediante Whole Slide Images enfrenta desafíos de interpretabilidad y actualización en sistemas de inteligencia artificial actuales. Este trabajo desarrolla un sistema multi-agente que integra Retrieval-Augmented Generation visual y textual para asistencia diagnóstica interpretable, combinando: (1) clasificación de parches mediante recuperación de casos similares, (2) respuestas conversacionales enriquecidas con guías clínicas, y (3) validación mediante arquitectura generador-evaluador. El clasificador ImageRAG alcanza 87.5% de accuracy global (99% en Normal, 95% en Invasivo) manteniendo interpretabilidad visual. La arquitectura multi-agente corrige errores de coherencia, pero requiere corpus textual abundante para precisión óptima. La latencia de clasificación WSI (3-8 min) es factible, aunque la conversacional (76-239s) necesita optimización. El mecanismo Human-in-the-Loop permite actualización incremental sin reentrenamiento, superando la rigidez de modelos end-to-end. Los resultados validan que sistemas basados en retrieval pueden ser clínicamente preferibles cuando interpretabilidad es crítica.

Multi-Agent Artificial Intelligence System for Precision

Oncology

Sebastián Andrés Donoso Arquero

Thesis submitted in partial fulfillment of the requirements for the degree of Electronic Engineering, specialization in Computer Structures and Systems, sub-specialization in Computer Science.

Universidad Técnica Federico Santa María

Thesis Advisor: Dr. Werner Creixell Fuentes

Co-Advisor: Dr. Alejandro Weinstein

MARZO 2026

Abstract

Breast cancer diagnosis through Whole Slide Images faces interpretability and update challenges in current artificial intelligence systems. This work develops a multi-agent system integrating visual and textual Retrieval-Augmented Generation for interpretable diagnostic assistance, combining: (1) patch classification through similar case retrieval, (2) conversational responses enriched with clinical guidelines, and (3) validation through generator-evaluator architecture. The ImageRAG classifier achieves 87.5% overall accuracy (99% on Normal, 95% on Invasive) while maintaining visual interpretability. The multi-agent architecture corrects coherence errors but requires abundant textual corpus for optimal precision. WSI classification latency (3-8 min) is feasible, though conversational latency (76-239s) requires optimization. The Human-in-the-Loop mechanism enables incremental updates without retraining, overcoming the rigidity of end-to-end models. Results validate that retrieval-based systems can be clinically preferable when interpretability is critical.

Glosario

API	Application Programming Interface. Conjunto de definiciones y protocolos para construir e integrar software, permitiendo comunicación entre componentes.
AUC	Area Under the Curve. Métrica de desempeño de clasificadores que representa el área bajo la curva ROC, valores cercanos a 1.0 indican excelente discriminación.
CAP	College of American Pathologists. Organización que establece estándares de reporte patológico mediante protocolos estructurados ampliamente adoptados en Estados Unidos.
CLIP	Contrastive Language-Image Pre-training. Modelo que aprende representaciones conjuntas de imágenes y texto mediante aprendizaje contrastivo en 400M pares.
CNN	Convolutional Neural Network. Arquitectura de red neuronal especializada en procesamiento de imágenes mediante capas convolucionales que extraen características jerárquicas.
CoT	Chain-of-Thought. Técnica de prompting que instruye a LLMs a explicitar su razonamiento paso a paso antes de generar respuesta final, mejorando desempeño en tareas complejas.
CPU	Central Processing Unit. Procesador principal del sistema que ejecuta instrucciones de software de propósito general.
CV	Cross-Validation. Técnica de validación que particiona datos en subconjuntos, entrenando y evaluando iterativamente para estimar generalización del modelo.
DCIS	Ductal Carcinoma In Situ. Proliferación neoplásica de células epite-

liales confinada al sistema ductal mamario sin invasión del estroma circundante.

FAISS Facebook AI Similarity Search. Biblioteca para búsqueda eficiente de vecinos más cercanos en espacios vectoriales de alta dimensión, con soporte para índices exactos y aproximados.

FastAPI Framework web moderno de Python para construcción de APIs, basado en estándares OpenAPI y diseñado para desarrollo rápido con validación automática de tipos.

FP8 Floating Point 8-bit. Formato numérico de precisión reducida (8 bits) soportado nativamente por GPUs modernas (Hopper, Ada Lovelace) para acelerar inferencia de LLMs con mínima degradación.

GPT Generative Pre-trained Transformer. Familia de modelos de lenguaje autoregresivos (decoder-only) entrenados en predicción de siguiente token.

GPU Graphics Processing Unit. Procesador especializado en cómputo paralelo masivo, esencial para entrenamiento e inferencia de modelos de deep learning.

H&E Hematoxilina y Eosina. Técnica de tinción histológica estándar donde hematoxilina tiñe núcleos celulares (azul-púrpura) y eosina tiñe citoplasma y tejido conectivo (rosa).

Human-in-the-Loop

Paradigma de diseño donde humanos intervienen en el ciclo de aprendizaje del sistema, proporcionando supervisión, correcciones o validación de decisiones automatizadas.

IHQ Inmunohistoquímica. Técnica que utiliza anticuerpos para detectar antígenos específicos en tejidos, usada para determinar expresión de

receptores hormonales y marcadores moleculares.

- JSON** JavaScript Object Notation. Formato ligero de intercambio de datos basado en texto, fácilmente legible por humanos y máquinas.
- K-NN** K-Nearest Neighbors. Algoritmo de clasificación no paramétrico que asigna etiquetas basándose en las k instancias más cercanas en un espacio de características.
- LLM** Large Language Model. Modelos de lenguaje de gran escala (típicamente $>1B$ parámetros) entrenados mediante predicción de siguiente token en corpus masivos, capaces de tareas de comprensión y generación de lenguaje natural.
- MCP** Model Context Protocol. Protocolo de comunicación estandarizado propuesto por Anthropic para interacción cliente-servidor entre aplicaciones y modelos de IA, permitiendo ejecución de herramientas arbitrarias.
- MIL** Multiple Instance Learning. Paradigma de aprendizaje supervisado débil donde las etiquetas se asignan a conjuntos (bolsas) de instancias en lugar de instancias individuales, común en análisis de WSI.
- NCCN** National Comprehensive Cancer Network. Alianza de centros oncológicos que publica guías de práctica clínica basadas en evidencia, actualizadas trimestralmente.
- NLP** Natural Language Processing. Procesamiento de Lenguaje Natural. Área de la inteligencia artificial enfocada en la interacción entre computadoras y lenguaje humano.
- PyTorch** Biblioteca de deep learning de código abierto desarrollada por Meta, con API flexible y diseño basado en define-by-run para construcción dinámica de grafos computacionales.



RAG	Retrieval-Augmented Generation. Arquitectura que combina recuperación de información de bases de conocimiento externas con generación de respuestas mediante modelos de lenguaje, mejorando precisión y reduciendo alucinaciones.
RAM	Random Access Memory. Memoria volátil de acceso aleatorio del sistema, usada para almacenamiento temporal de datos y programas en ejecución.
REST	Representational State Transfer. Estilo arquitectural para diseño de APIs web basado en protocolo HTTP con operaciones CRUD estándar.
ROC	Receiver Operating Characteristic. Curva que grafica tasa de verdaderos positivos vs tasa de falsos positivos a diferentes umbrales de clasificación.
SVS	Aperio ScanScope Virtual Slide. Formato propietario de Leica Biosystems para almacenamiento de WSI con estructura piramidal multinivel y compresión JPEG.
ViT	Vision Transformer. Arquitectura basada en mecanismos de atención para procesamiento de imágenes, alternativa a CNNs que captura dependencias globales.
VRAM	Video Random Access Memory. Memoria dedicada de la GPU para almacenamiento de modelos, activaciones y datos durante cómputo.
WHO	World Health Organization. Organización Mundial de la Salud. Publica sistemas de clasificación internacional de tumores, incluyendo el WHO Classification of Tumours.
WSI	Whole Slide Image. Imagen digital de alta resolución (típicamente gigapíxel) de un tejido, escaneada a magnificaciones de 20× o 40×.

Índice de contenidos

Glosario	iii
1 Introducción	1
1.1 Solución Propuesta	3
1.2 Objetivos	4
1.3 Organización del Documento	4
2 Estado del Arte	6
2.1 Inteligencia Artificial en Imágenes Médicas	6
2.1.1 Deep Learning en Patología Digital	8
2.1.2 Dificultades presentes en patología	10
2.2 Patología Digital y Whole Slide Images (WSI)	12
2.3 Retrieval-Augmented Generation (RAG)	14
2.3.1 RAG Textual	17
2.3.2 RAG Visual	18
2.3.3 Métricas de Similitud y K-NN	20
2.4 Large Language Models (LLMs) en Medicina	20
2.4.1 Evolución	21
2.5 Sistemas de Apoyo Diagnóstico en Patología Mamaria	22
2.5.1 Clasificación Automática de Tejido Mamario	22
2.5.2 Integración de RAG y LLM en Aplicaciones Médicas	23
2.6 Análisis Comparativo de Trabajos Relacionados	24

2.6.1	Gaps de Investigación	24
2.6.2	Posicionamiento de Este Trabajo	25
3	Diseño e Implementación de la Solución	27
3.1	Ambiente de Desarrollo e Infraestructura	27
3.1.1	Dependencias de Software	27
3.1.2	Hardware de Procesamiento	28
3.2	Adquisición y Preprocesamiento de Datos	29
3.3	Visión General del Sistema	31
3.4	Diseño del Módulo Pre Procesamiento de WSI	34
3.5	Diseño del Módulo de Recuperación Visual (Image RAG)	37
3.5.1	Arquitectura del Sistema RAG Visual	38
3.5.1.1	Fase 1: Construcción del Índice Vectorial	38
3.5.1.2	Fase 2: Búsqueda por Similitud (Inference)	42
3.5.2	Manejo de Desbalanceo de Clases	46
3.5.3	Limitaciones del Diseño RAG	46
3.6	Diseño del Módulo Conversacional (Agentes LLM)	47
3.6.1	Motivación del Diseño Multi-Agentes	47
3.6.2	Componentes de la Arquitectura	48
3.6.2.1	Agente Generador (Small Language Model):	48
3.6.2.2	Agente Evaluador (MedGemma 27B):	49
3.6.3	Estrategia de Prompting	50
3.6.3.1	Prompt del Generador	51
3.6.3.2	Prompt del Evaluador	52
3.6.4	Enriquecimiento Contextual mediante RAG Textual	52
3.6.4.1	Construcción de la Base de Conocimientos	52
3.6.4.2	Manejo de Historial	56
3.6.5	Consultas Multimodales sobre Parches Específicos	56
3.6.6	Limitaciones del Diseño de Agentes LLM	57

3.7	Interfaz de Usuario	58
3.7.1	Flujo De Trabajo Human-In-The-Loop	59
3.7.2	Visualización de Parches Similares ImageRag	62
3.7.3	Comunicación entre Módulos	63
3.7.4	Despliegue	66
3.7.5	Limitaciones de la Interfaz de Usuario	66
4	Resultados y Análisis	67
4.1	Definición de Pruebas y Métricas	67
4.1.1	Rendimiento del Clasificador (Image RAG)	67
4.1.2	Eficacia de la Orquestación LLM	68
4.1.3	Latencia del Sistema	71
4.2	Resultados Experimentales	71
4.2.1	Desempeño del Clasificador Image RAG	71
4.2.1.1	Análisis de la Matriz de Confusión	73
4.2.2	Evaluación de la Orquestación LLM	76
4.3	Rendimiento del Sistema y Validación Técnica	87
4.3.1	Clasificación de una WSI usando ImageRAG para diferentes tamaños de parches	87
4.3.2	Recuperación de parches similares (múltiples vecinos)	88
4.3.3	Generación de respuesta de agentes LLM (Múltiples modelos)	89
5	Conclusiones y Trabajo Futuro	91
5.1	Implementación de Sistema Multiagente para Asistir Análisis De Imagen de Tejido WSI (Whole Slide Image)	91
5.2	Recaudar Documentos e Imágenes Para Construcción de Sistema Retrieval Augmented Generation(RAG)	92
5.3	Evaluación de Distintos Modelos SML (Small Language Model).	93
5.4	Limitaciones	93
5.5	Trabajo Futuro	94



Referencias	95
Apéndice	103

Lista de Figuras

1.1	Comparación visual entre arquitectura CNN y arquitectura VIT. en (a) observamos la arquitectura VIT, donde destaca su mecanismo de multi-atención. En (b) tenemos el proceso de Convolution asociada a las CNNS. Figura obtenida del survey [18].	7
1.2	Muestras tintadas con H&E obtenidas por diferentes escáneres. Los diferentes escáneres representan los colores de diferente intensidad. . . .	11
2.3	Estructura piramidal de una WSI en patología digital. La figura ha sido extraída desde [35]	13
3.4	Arquitectura RAG, extraída desde trabajo original [23]. Dado una query, esta se codifica (embedding) y se recuperan los documentos similares desde un índice. Los documentos recuperados son entregados como contexto al generador encargado de responder la query.	15
3.1	Arquitectura general del sistema de apoyo. El flujo comprende cuatro módulos principales: preprocesamiento de WSI, clasificación mediante ImageRAG, asistente conversacional LLM, e interfaz de interacción usuario-sistema con mecanismo de corrección	31
4.2	Arquitectura del pipeline de pre-procesamiento. El sub-módulo de TIA-ToolBox se encarga de leer la WSI, aplicar segmentación Otsu y recortar los parches, mientras que el filtrado heurístico y la generación de embeddings fueron incorporaciones externas.	34

5.3	Población inicial del ImageRAG con dataset ICIAR 2018. En batches de 32 imágenes, se calcula su embedding, almacenan los datos de ubicación y categoría de cada imagen, y finalmente se insertan en el índice FAISS.	39
5.4	Recuperación y clasificación de parches con ImageRAG. Dado un parche, obtenemos su embedding con MedSigLIP y a través de agregación ponderada obtenemos clasificación (que corresponde a la categoría con mayor puntaje).	42
6.5	Arquitectura del sistema de chat con evaluación en cascada. El flujo incluye recuperación de contexto (RAG textual + estadísticas WSI), generación inicial por modelo ligero y evaluación por modelo especializado (modelo que reflexiona a través del modo Thinking).	47
6.6	Esquema de población de RAG textual.	54
6.7	Recuperación de documentos desde base de datos vectorial. Los Top-K Mejores embeddings se concatenan al prompt del modelo generador, proporcionándole el contexto adecuado.	55
7.8	El usuario observa WSI, presiona un parche, lo corrige, y se almacena en feedback para eventual mejora del modelo.	60
7.9	Interfaz de corrección de clasificación. El usuario al seleccionar un parche, tiene la posibilidad de modificar la clasificación de este, lo que provoca que el JSON con la información asociada a la WSI se actualice. Eventualmente, podemos extraer este parche y su nueva clasificación e incorporarlas al RAG visual.	60
7.10	Interfaz de usuario del sistema. A la derecha se observa la ventana para interactuar con los agentes, en la parte inferior se observan las estadísticas de clasificación y en la parte central se observa la WSI clasificada.	62
7.11	Visualización de imágenes similares al parche seleccionado. Esto permite al patólogo analizar la razón de la elección de categoría.	63

7.12 Diagrama sobre el funcionamiento del protocolo MCP. El módulo «MCP» orquesta las llamadas a las distintas herramientas. Las herramientas co- rresponden a rutinas de inferencia de los small model languages.	65
2.1 Matriz de confusión del Image RAG.	73
2.2 Resultados k-fold cross validation para clasificador Image RAG.	76
5.1 Whole Slide Image utilizada para prueba de latencia en clasificación. Esta muestra fue proporcionada por Environ SPA.	104

Introducción

El cáncer de mama es una enfermedad manifestada por el crecimiento anormal y desordenado de las células que revisten el epitelio de los conductos o lóbulos mamarios, y que poseen la capacidad de diseminarse. En el año 2020, la International Agency of Research on Cancer estimó que se produjeron cerca de 2.260.000 nuevos casos y 1.790.000 personas fallecieron por esta causa. En Chile, ese mismo año se estimó que el cáncer de mama constituyó el 10% de todos los casos nuevos de cáncer en la población general, y se produjeron 1657 decesos, en donde 13 correspondieron a población masculina [27]. Por lo mismo, el cáncer de mama en Chile ocupa el primer lugar en mortalidad por cáncer en la mujer, transformándose en un problema de salud pública de gran impacto. Para combatir el cáncer de mama, es necesario un esfuerzo en conjunto entre las entidades estatales, la academia y las sociedades científicas, de manera que se puedan levantar estrategias más efectivas para combatir esta enfermedad.

La tinción con Hematoxilina y Eosina (H&E) constituye la técnica histopatológica estándar para diagnóstico de cáncer de mama, permitiendo la visualización de estructuras celulares mediante contraste cromático: núcleos celulares en tonos azul-púrpura (hematoxilina) y citoplasma en rosa (eosina). A pesar de su adopción universal, esta técnica presenta variabilidad: cada patólogo ajusta concentraciones de tintes y tiempos de exposición según preferencia personal, generando inconsistencias que afectan la reproducibilidad del diagnóstico.

El diagnóstico de cáncer de mama mediante H&E requiere evaluación sistemática de múltiples criterios morfológicos (como por ejemplo, la estructura microscópica de un tejido), asignando puntajes que determinan decisiones en la terapia. Este proceso, subjetivo, depende de la experiencia y criterio del patólogo, resultando en diferencias entre profesionales.

La digitalización de láminas histológicas ha transformado el flujo de trabajo patológico: escáneres automatizados generan Whole Slide Images (WSI) de resolución gigapíxel (típicamente $50,000 \times 50,000$ píxeles a magnificación $20\times$), permitiendo visualización en pantalla sin necesidad de microscopios. Esta transición habilita nuevas capacidades: almacenamiento digital permanente, consulta remota entre instituciones, análisis cuantitativo automatizado y, fundamentalmente, aplicación de algoritmos de inteligencia artificial sobre datos digitalizados.

La aplicación de inteligencia artificial al diagnóstico histopatológico enfrenta tres desafíos fundamentales:

Escasez de Datos Etiquetados de Alta Calidad

El entrenamiento de modelos de aprendizaje profundo requiere datasets extensos con anotaciones precisas. En histopatología, esto implica obtener datos de pacientes protegidos por regulaciones de privacidad, que se encuentren detalladamente anotados (esto requiere horas de trabajo por parte de un experto) y que se encuentren estandarizados, ya que cada patólogo puede entregar un diagnóstico diferente, introduciendo ruido en el dataset.

Aunque existen datasets públicos, ninguno proporciona la granularidad diagnóstica completa necesaria para replicar el flujo de trabajo patológico real.

Falta de Interpretabilidad en Modelos End-to-End

Los sistemas de clasificación automática basados en redes neuronales convolucionales (CNNs) o Vision Transformers (ViTs) han alcanzado accuracies superiores a 90%

en benchmarks específicos [43]. Sin embargo, operan como "cajas negras": generan predicciones categóricas (e.g., "carcinoma invasivo") sin explicación auditable del razonamiento diagnóstico.

Rigidez Post-Despliegue y Ausencia de Actualización Dinámica

Modelos entrenados mediante fine-tuning de arquitecturas fundacionales (e.g., ViT-Large con $>300M$ parámetros) son estáticos tras despliegue: incorporar nuevo conocimiento (actualización de guías clínicas, descubrimiento de biomarcadores, casos raros) requiere reentrenamiento completo con infraestructura GPU especializada, demandando días-semanas de cómputo y expertise técnico.

Esta rigidez es incompatible con la naturaleza dinámica del conocimiento médico: guías de práctica clínica se actualizan constantemente, nueva evidencia se publica continuamente, y los laboratorios pueden contar con nuevas muestras con el tiempo.

1.1 Solución Propuesta

Este trabajo propone un sistema de asistencia diagnóstica basado en **Retrieval-Augmented Generation (RAG)** dual que combina recuperación de evidencia visual y textual con orquestación de modelos LLMs, priorizando interpretabilidad y actualizabilidad.

Enfoque RAG Dual

A diferencia de clasificadores end-to-end que memorizan patrones en pesos neuronales, el sistema propuesto opera mediante recuperación explícita de información.

Un sistema RAG se puede pensar como una base de datos que contiene la representación vectorial de imágenes (RAG visual) o texto (RAG textual). Esto nos permite comparar imágenes o textos nuevos con las representaciones ya existentes, a través de operaciones matemáticas de bajo costo computacional. Gracias a la naturaleza del RAG, podemos entregar no sólo una clasificación, sino que también evidencia visual/escrita de las razones que fundamentan la decisión del modelo.

Arquitectura Multi-Agente Generador-Evaluador

Para equilibrar latencia y calidad de respuesta, se implementa orquestación de dos niveles (1) Modelo generador pequeño SLM entre 4B-8B de parámetros, que se encarga de entregar una respuesta rápida basándose en el contexto recuperado del RAG y la clasificación de la WSI, y (2) Modelo evaluador más complejo de 27B de parámetros, encargado de verificar que la respuesta entregada por el modelo generador cumpla estándares basándose en una pauta.

Interactividad Human-in-the-Loop

Finalmente, la interfaz web implementa ciclo de retroalimentación continua: patólogos pueden corregir clasificaciones erróneas de parches, que eventualmente se pueden incorporar al RAG visual para mejorar el sistema. Esto contrasta con modelos estáticos que requieren reentrenamiento completo para incorporar correcciones.

1.2 Objetivos

1. Implementar sistema multiagente para asistir análisis de imagen de tejido WSI (Whole Slide Image).
2. Recaudar documentos e imágenes para construcción de sistema Retrieval Augmented Generation (RAG).
3. Evaluación de distintos modelos SML (Small Language Model).

1.3 Organización del Documento

El resto de este documento se estructura como sigue:

- **Capítulo 2 - Estado del Arte:** Revisión de literatura en IA aplicada a imagenología médica, patología digital, sistemas RAG y LLMs médicos. Identificación de gaps que motivan este trabajo.

- **Capítulo 3 - Diseño e Implementación de la Solución:** Diseño conceptual del sistema propuesto, justificación de decisiones arquitecturales (RAG vs fine-tuning, generador-evaluador, selección de modelos) e implementación de la prueba de concepto.
- **Capítulo 4 - Resultados y Análisis:** Evaluación experimental del clasificador Image RAG (accuracy, matriz de confusión), evaluación cualitativa de respuestas LLM, medición de latencia computacional.
- **Capítulo 5 - Conclusiones y Trabajo Futuro:** Síntesis de hallazgos, cumplimiento de objetivos, limitaciones identificadas y recomendaciones para evolución del sistema.

Estado del Arte

Este capítulo presenta una revisión sistemática del estado actual de las tecnologías que sustentan esta investigación. Se estructura en secciones que abarcan desde los fundamentos de IA en medicina hasta sistemas específicos de apoyo diagnóstico en patología mamaria. El análisis identifica las fortalezas de los enfoques existentes, así como las limitaciones que motivan el desarrollo de este trabajo.

2.1 Inteligencia Artificial en Imágenes Médicas

La aplicación de inteligencia artificial (IA) en imagenología médica ha experimentado una transformación en las últimas dos décadas, evolucionando desde sistemas basados en reglas heurísticas hasta arquitecturas de aprendizaje profundo capaces de igualar el rendimiento humano en tareas diagnósticas específicas [9].

Los sistemas de Diagnóstico Asistido por Computador (Computer-Aided Diagnosis, CAD) representan una de las primeras aplicaciones exitosas de IA en medicina clínica. Su desarrollo puede dividirse en tres generaciones tecnológicas claramente diferenciadas.

Los primeros sistemas CAD se fundamentaron en la extracción manual de características (*hand-crafted features*) y reglas de decisión definidas por expertos [17]. En el contexto de imágenes mamarias, pioneros como Giger et al. (1994) desarrollaron algoritmos para detectar microcalcificaciones en mamografías mediante análisis de textura

y umbralización adaptativa [31].

Luego, la disponibilidad de datasets digitalizados y el avance en algoritmos de aprendizaje automático permitieron una transición hacia métodos basados en datos. Técnicas como Random Forests, Gradient Boosting y Support Vector Machines (SVM) con kernels no lineales reemplazaron las reglas expertas [13].

Un hito representativo fue el trabajo de Cruz-Roa et al. (2014) en detección automática de carcinoma invasivo en histopatología de mama, logrando AUC de 0.96 mediante Random Forests con features de textura Haralick [6]. No obstante, el cuello de botella persistía: la obtención de features requería expertise del dominio y no escalaba a modalidades nuevas sin rediseño sustancial.

En nuestra época, la aparición de redes neuronales convolucionales (CNNs) marcó un punto de inflexión. El trabajo seminal de Krizhevsky et al. (2012) con AlexNet en ImageNet demostró que arquitecturas profundas podían aprender representaciones jerárquicas directamente desde píxeles, eliminando la necesidad de features manuales [21]. Sin embargo, soluciones basadas en transformers son los estándares a día de hoy. En la figura 1.1

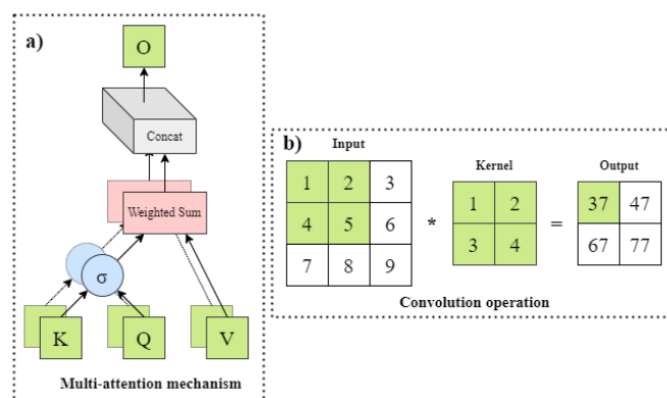


Figure 1.1: Comparación visual entre arquitectura CNN y arquitectura ViT. en (a) observamos la arquitectura ViT, donde destaca su mecanismo de multi-atención. En (b) tenemos el proceso de Convolution asociada a las CNNs. Figura obtenida del survey [18].

2.1.1 Deep Learning en Patología Digital

La patología digital, definida como la adquisición, gestión y visualización de Whole Slide Images (WSI), presenta desafíos computacionales únicos que la distinguen de otras modalidades de imágenes médicas [1].

A diferencia de radiografías (resolución típica $\sim 2K \times 2K$ píxeles) o resonancias magnéticas (volúmenes de $\sim 512 \times 512 \times 200$), las WSI operan en una escala de **gigapíxeles**:

- **Dimensiones Típicas:** $50,000 \times 50,000$ a $200,000 \times 200,000$ píxeles (escaneo a $40\times$).
- **Tamaño de Archivo:** 1-5 GB por imagen sin comprimir; 200-500 MB con compresión JPEG2000.
- **Información Multi-Escala:** Estructuras diagnósticas relevantes abarcan desde núcleos celulares ($5-10 \mu\text{m}$, visibles a $40\times$) hasta arquitectura tisular global (varios milímetros, observable a $2.5\times$).

Esta escala inhabilita el procesamiento directo: una WSI de $100,000 \times 100,000$ píxeles requiere 30 GB de RAM solo para cargarse sin compresión (asumiendo RGB de 8 bits). En consecuencia, los enfoques de deep learning en patología adoptan estrategias de **análisis basado en parches** (*patch-based analysis*) [45].

La literatura identifica tres paradigmas principales para aplicar deep learning a las WSI:

Clasificación Débilmente Supervisada (Weakly-Supervised Learning)

Dado que anotar píxel por píxel en WSI gigapíxel es difícil, la mayoría de datasets solo proporcionan etiquetas a nivel de imagen completa (e.g., "tumor presente/ausente"). El desafío es entrenar modelos que localicen regiones discriminativas sin supervisión explícita.

Multiple Instance Learning (MIL): Trata cada WSI como una "bolsa" de parches. La etiqueta de la bolsa es positiva si al menos un parche contiene la característica de

interés. Ilse et al. (2018) propusieron Attention-based MIL, donde un mecanismo de atención aprende a ponderar parches relevantes [15]. Esta arquitectura logró AUC de 0.92 en clasificación de subtipos de cáncer de pulmón en TCGA, usando solo etiquetas de slide-level.

Clasificación Fuertemente Supervisada con Segmentación

Cuando se dispone de anotaciones precisas (típicamente para desafíos de investigación), se entrenan modelos de segmentación semántica:

- **U-Net y Variantes:** Arquitectura encoder-decoder con skip connections, estándar para segmentación médica.
- **Mask R-CNN:** Para detección de instancias (núcleos celulares individuales).

Aprendizaje Auto-Supervisado (Self-Supervised Learning)

Ante la escasez de datos anotados, métodos de pre-entrenamiento auto-supervisado extraen representaciones útiles de datos sin etiquetar:

- **Contrastive Learning:** SimCLR, MoCo – aprenden embeddings invariantes mediante contrastes positivos/negativos.
- **Masked Image Modeling:** Inspirado por BERT en NLP, MAE (Masked Autoencoder) para imágenes aprende a reconstruir regiones enmascaradas..

Modelos Fundacionales en Patología

Recientemente, la comunidad ha adoptado el concepto de **modelos fundacionales** (*foundation models*): redes masivas pre-entrenadas en corpus gigantes que se adaptan a tareas downstream mediante fine-tuning mínimo.

Ejemplos representativos:

- **CTransPath (2021):** ViT-Small pre-entrenado en 15 millones de parches de 32 tipos de cáncer mediante contrastive learning. Superó ResNet-50 en 7 de 9 datasets de clasificación tumoral [43].

- **UNI (2024):** Encoder DINOv2 pre-entrenado en 100 millones de parches de más de 100,000 WSI de 20 instituciones. Alcanzó desempeño state-of-the-art en 9 benchmarks sin fine-tuning, solo con clasificadores lineales [5].
- **Prov-GigaPath (2024):** Modelo visual-lingüístico con 1 billón de parámetros, entrenado en 1.3 billones de parches con descripciones textuales de reportes patológicos. Primera demostración de zero-shot reasoning en patología (e.g., "encuentra regiones con alta densidad linfocitaria") [47].
- **MUSK (2024):** Modelo multimodal que utiliza transformer con Unified maSKed modeling (MUSK), pre entrenado con más de 50 millones de imágenes patológicas de 11577 pacientes y un billón de texto patológico usando unified masked modelling. Con mínimo o sin entrenamiento, MUSK demuestra un desempeño superior en 23 patch-level y slide-level benchmarks, incluyendo image-to-text, recuperación text-to-image, visual question answering, clasificación de imágenes y predicción de biomarcadores moleculares [46].

2.1.2 Dificultades presentes en patología

A pesar de los avances, persisten desafíos fundamentales que limitan la adopción clínica de sistemas de IA en patología.

Variabilidad Inter-Institucional

La histopatología carece de estandarización comparable a radiología. Factores como:

- Diferencias en protocolos de tinción H&E (concentración de hematoxilina, tiempo de exposición).
- Variabilidad entre escáneres (Aperio, Hamamatsu, Leica) en balance de color y compresión.

Estos factores introducen **domain shift**: modelos entrenados en Institución A degradan significativamente cuando se aplican en Institución B sin re-calibración [41]. Un ejemplo

de esto se puede ver en la figura [1.2](#)

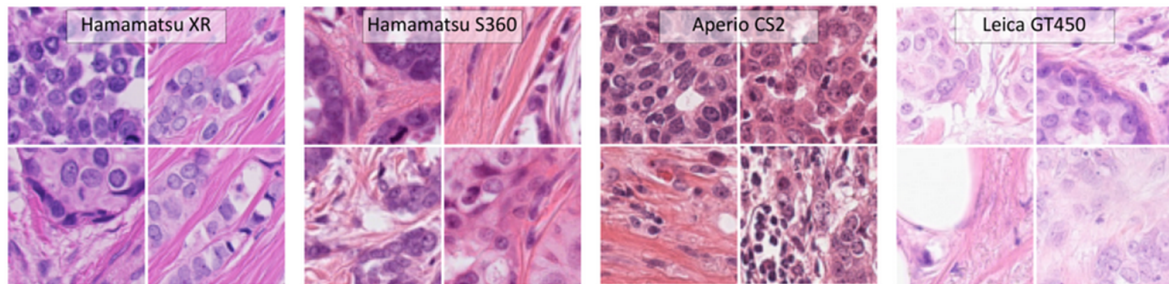


Figure 1.2: Muestras tintadas con H&E obtenidas por diferentes escáneres. Los diferentes escáneres representan los colores de diferente intensidad.

Escasez de Datos Anotados de Alta Calidad

A diferencia de ImageNet (14M imágenes etiquetadas), los datasets médicos son órdenes de magnitud menores:

- CAMELYON16: 400 WSI con anotaciones de metástasis.
- TCGA: 11,000 WSI con etiquetas de subtipo molecular, pero sin segmentación.
- ICIAR 2018 (usado en este trabajo): **solo 400 imágenes**.

Esta escasez limita la capacidad de generalización de modelos supervisados.

Interpretabilidad y Explicabilidad

Los modelos de deep learning operan como "cajas negras", lo cual es inaceptable en medicina donde las decisiones deben ser auditables y justificables legalmente. Un patólogo no puede firmar un reporte basándose en una predicción sin comprender su razonamiento.

Enfoques de explicabilidad:

- **Mapas de Atención:** Visualizar qué regiones el modelo considera importantes (Grad-CAM, Attention Rollout) [\[39\]](#).

- **Retrieval-Based Explanations:** Mostrar casos similares del dataset de entrenamiento (enfoque adoptado en esta memoria mediante RAG visual).

La siguiente sección examina los fundamentos de patología digital y las características técnicas de las Whole Slide Images que motivan el diseño del sistema propuesto.

2.2 Patología Digital y Whole Slide Images (WSI)

La patología digital representa la transición de la microscopía óptica tradicional hacia el análisis completamente digital [10]. Esta transformación tecnológica no solo modifica el flujo de trabajo, sino que habilita nuevas capacidades analíticas imposibles de forma analógica, incluyendo la integración de sistemas de inteligencia artificial.

Digitalización de Tejidos

Los escáneres modernos operan en magnificaciones de 20x o 40x (equivalente a una resolución de ~ 0.5 o $\sim 0.25 \mu\text{m}/\text{píxel}$), generando imágenes con dimensiones típicas de 50000 x 50000 a 200000 x 200000 píxeles. Para optimizar almacenamiento y visualización, las WSI se codifican en **formato piramidal multi-resolución**, donde la imagen original se acompaña de versiones progresivamente reducidas (downsampled) a factores de 2x, 4x, 8x, etc. Esta estructura permite navegación eficiente: al hacer zoom out, el visualizador carga niveles de baja resolución; al magnificar regiones específicas, se accede al nivel de máxima resolución solo para el área visible. La figura 2.3 representa la estructura de los archivos que contienen las WSI.

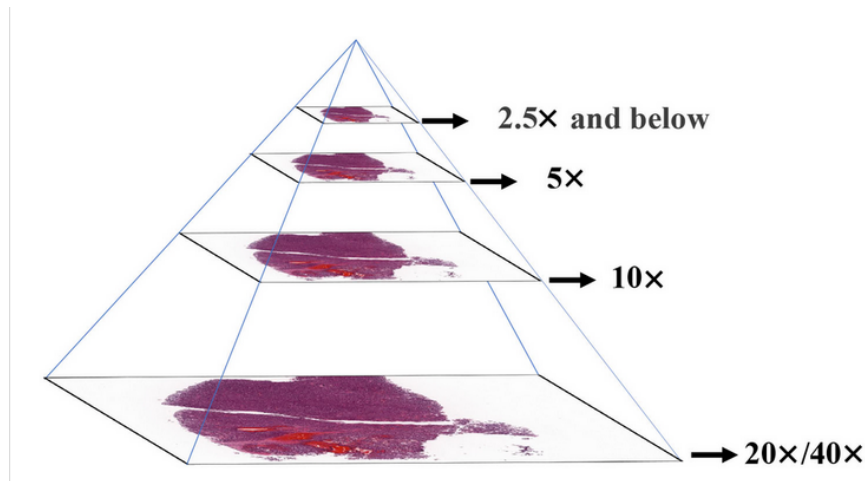


Figure 2.3: Estructura piramidal de una WSI en patología digital. La figura ha sido extraída desde [35]

Características Técnicas de las WSI

Las WSI se distinguen por su **resolución gigapíxel**, varios órdenes de magnitud superior a modalidades radiológicas tradicionales:

Table 2.1: Comparación de dimensiones entre modalidades de imágenes médicas.

Modalidad	Resolución Típica	Tamaño Archivo	Píxeles Totales
Radiografía	2048 × 2048	8 MB	4 MP
TC (1 corte)	512 × 512	0.5 MB	0.26 MP
RM (volumen)	256 × 256 × 180	24 MB	12 MP
WSI (40×)	100,000 × 100,000	1-3 GB	10,000 MP

Esta escala impone restricciones computacionales críticas.

Dado que incluso GPUs de alto rendimiento (e.g., NVIDIA A100 con 80 GB) no pueden procesar imágenes completas, todos los algoritmos de análisis adoptan estrategias de **procesamiento basado en parches**.

Desafíos Computacionales: Procesamiento de GigaPíxeles

A diferencia de modalidades con protocolos estandarizados (e.g., TC/RM con parámetros de adquisición certificados), la histopatología presenta variabilidad sustancial

en:

- **Tinción:** Concentración de hematoxilina, tiempo de exposición a eosina, temperatura de procesamiento. Resulta en espectros de color inconsistentes entre instituciones.
- **Grosor de Sección:** Variaciones de 4-6 μm afectan densidad celular aparente.
- **Escáneres:** Diferentes fabricantes emplean sensores, ópticas y algoritmos de compresión distintos, introduciendo artefactos específicos del dispositivo.

Esta variabilidad es la causa principal de **domain shift** en modelos de deep learning: un clasificador entrenado en datos de Institución A puede degradar 10-15% en accuracy al aplicarse en Institución B sin adaptación [41]. Técnicas de normalización de color (e.g., Macenko) son esenciales para robustez cross-institutional [26].

2.3 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) representa un cambio en el diseño de sistemas de inteligencia artificial. Alejándose de modelos puros hacia arquitecturas híbridas que combinan recuperación explícita de información con generación de texto [23]. Esta sección examina los fundamentos teóricos de RAG, su implementación en dominios textuales y visuales, y sus adaptaciones específicas para el contexto médico.

Fundamentos RAG

La figura 3.4 corresponde a un enfoque resumido de la arquitectura RAG:

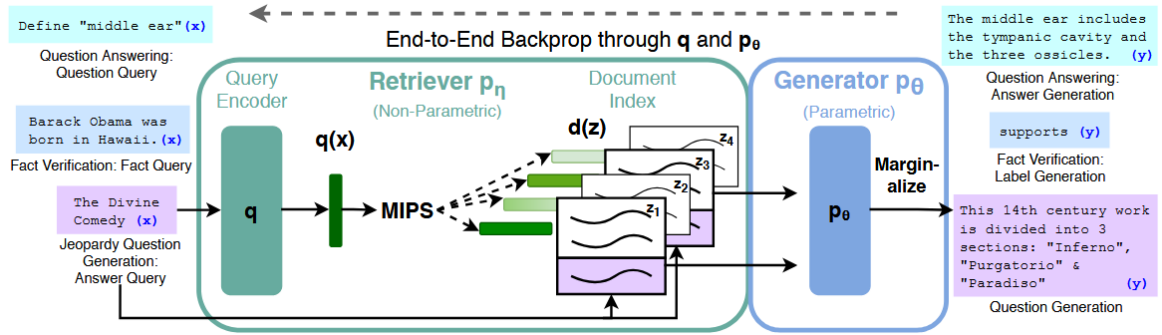


Figure 3.4: Arquitectura RAG, extraída desde trabajo original [23]. Dado una query, esta se codifica (embedding) y se recuperan los documentos similares desde un índice. Los documentos recuperados son entregados como contexto al generador encargado de responder la query.

El paradigma RAG descompone el proceso de generación de respuestas en dos fases diferenciadas:

1. **Retrieval (Recuperación):** Dado un query q , se recuperan los k documentos o ejemplos más relevantes $D = \{d_1, d_2, \dots, d_k\}$ desde una base de conocimiento externa \mathcal{K} mediante búsqueda por similitud:

$$D = \text{top-}k(\mathcal{K}, q) = \underset{d \in \mathcal{K}}{\text{arg max}}_k \text{sim}(f(q), f(d)) \quad (3.1)$$

donde $f(\cdot)$ es una función de embedding que mapea queries y documentos a un espacio vectorial compartido, y $\text{sim}(\cdot, \cdot)$ es una métrica de similitud (típicamente similitud coseno o producto interno).

2. **Generation (Generación):** El conjunto recuperado D se concatena como contexto adicional al prompt del modelo generativo (típicamente un LLM), que produce la respuesta final:

$$p(y|q) = \sum_{d \in D} p(d|q) \cdot p(y|q, d) \quad (3.2)$$

donde $p(d|q)$ representa la relevancia del documento recuperado y $p(y|q, d)$ la

probabilidad de generar la respuesta y condicionada en el query y el contexto recuperado.

El trabajo seminal de Lewis et al. (2020) demostró que esta arquitectura supera modelos generativos puros (BART, T5) en tareas de question-answering intensivas en conocimiento, logrando 44.5% de exact match en Natural Questions vs 34.5% de BART sin retrieval [23].

Ventajas sobre Fine-tuning

RAG presenta ventajas estructurales sobre el tradicional fine-tuning (ver tabla 3.2) completo de modelos de lenguaje (LLM):

Table 3.2: Comparación RAG vs Fine-tuning en sistemas de conocimiento especializado.

Aspecto	Fine-tuning	RAG
Actualización	Reentrenamiento completo (costoso)	Modificación del índice (inmediata)
Interpretabilidad	Caja negra	Citas explícitas a fuentes
Hallucinations	Alto riesgo (conocimiento en pesos)	Reducido (grounding en evidencia)
Escalabilidad	Limitada por tamaño del modelo	Escalable con la base de datos

En medicina, la ventaja de actualización dinámica es crítica: guías clínicas evolucionan continuamente (e.g., NCCN actualiza protocolos trimestralmente). RAG permite incorporar nueva evidencia agregando documentos al índice sin reentrenamiento, mientras que fine-tuning requiere ciclos de entrenamiento de días-semanas con GPUs especializadas [11].

2.3.1 RAG Textual

La calidad del retrieval depende críticamente de la representación vectorial de texto. La evolución de modelos de embeddings puede trazarse en tres generaciones:

1. Embeddings Estáticos (Word2Vecm GloVe):

Asignan vectores fijos a palabras sin contexto. Limitación fundamental: "banco" (institución financiera) y "banco" (asiento) reciben el mismo vector [28].

2. Embeddings Contextuales (BERT, ELMo):

Generan representaciones dependientes del contexto circundante. BERT (2018) revolucionó NLP logrando state-of-the-art en 11 benchmarks mediante pre-entrenamiento bidireccional en masked language modeling [7].

3. Embeddings Especializados para Retrieval:

Sentence-BERT (SBERT) [36] modifica BERT para generar embeddings de oraciones enteras optimizados para similitud semántica.

Bases de Datos Vectoriales

El almacenamiento y búsqueda eficiente de embeddings a escala requiere estructuras de datos especializados. FAISS (Facebook AI Similarity Search)[16] es el estándar de facto, implementando múltiples variantes de índices (ver tabla 3.3).

Table 3.3: Principales tipos de índices FAISS y sus trade-offs.

Índice	Complejidad	Precisión	Uso Recomendado
IndexFlatL2	$O(nd)$	100%	Datasets pequeños (<100K)
IndexIVFFlat	$O(\sqrt{n} \cdot d)$	95-99%	Balance precisión-velocidad
IndexHNSW	$O(\log n)$	97-99%	Latencia crítica
IndexPQ	$O(n)$	85-95%	Memoria limitada

Alternativas comerciales como Pinecone, Weviate y ChromaDB abstraen la configuración de índices mediante APIs gestionadas, pero introducen dependencias de servicios externos y costos operacionales. Para esta prueba de concepto, se seleccionó IndexFlatL2 priorizando precisión máxima dado el tamaño manejable del dataset usado para esta prueba de concepto (ICIAR 400 imágenes [3]).

Estrategias de Retrieval

Más allá de recuperación simple top- k , la investigación ha desarrollado estrategias sofisticadas:

- **Dense Retrieval:** Embedding densos aprendidos (SBERT, E5). Superior en similitud semántica profunda pero computacionalmente intensivo.
- **Sparse Retrieval:** TF-IDF, BM25. Rápidos, pero limitados a matching léxico exacto.
- **Hybrid Retrieval:** Combina Dense Retrieval con Sparse Retrieval mediante rank fusion (e.g., Reciprocal Rank Fusion). Khattab et al. (2020) demostraron mejoras de 3-7% en MS MARCO con ColBERT, un modelo híbrido [19].
- **Re-ranking:** Tras recuperación inicial de 100 candidatos, un modelo cross-encoder (más costoso pero preciso) re-rankea los top-20 finales.

2.3.2 RAG Visual

La extensión de RAG al dominio visual requiere encoders que mapeen imágenes a espacios semánticos compartidos. CLIP (Contrastive Language-Image Pre-training) [34] estableció el paradigma dominante:

Arquitectura: Dos encoders (ViT para imagen, transformer para texto) entrenados conjuntamente en 400M pares imagen-texto de internet mediante contrastive loss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \left[\log \frac{\exp(\text{sim}(I_i, T_i)/\tau)}{\sum_j \exp(\text{sim}(I_i, T_j)/\tau)} \right] \quad (3.3)$$

donde I_i, T_i son embeddings de imagen y texto, τ es temperatura, y el denominador suma sobre todos los textos en el batch (negatives implícitos).

Capacidades: CLIP logra zero-shot classification competitiva con modelos supervisados en ImageNet (76.2% top-1 accuracy sin ver etiquetas de ImageNet) [34]. Su poder reside en la alineación visión-lenguaje: puede buscar imágenes mediante descripciones textuales arbitrarias sin fine-tuning.

Limitaciones: CLIP está entrenado en datos web generales (fotografías, arte, etc.) Su rendimiento degrada en dominios especializados como imágenes médicas, donde la terminología es técnica y las características visuales difieren radicalmente de imágenes naturales.

Evolución - SigLIP: Zhai et al. (2023) propusieron SigLIP (Sigmoid Loss for Language-Image Pre-training) [48], reemplazando softmax por sigmoid en la función de pérdida, permitiendo entrenar con batches más grandes y mejorando eficiencia:

$$\mathcal{L}_{\text{SigLIP}} = - \sum_{i,j} \log \sigma(y_{ij} \cdot \text{sim}(I_i, T_j)) \quad (3.4)$$

donde $y_{ij} \in \{-1, +1\}$ indica si (I_i, T_j) es par positivo o negativo. SigLIP-B/16 alcanza 82.9% en ImageNet, superando CLIP con menos datos.

Adaptaciones al Dominio Médico

La diferencia entre imágenes naturales y médicas motivó desarrollo de modelos especializados:

- **MedCLIP (Stanford, 2022):** Fine-tuning de CLIP en 217K pares radiografía-reporte de MIMIC-CXR. Mejora zero-shot classification en 8 datasets radiológicos, promedio 4.6% sobre CLIP base [44].
- **BiomedCLIP (Microsoft, 2023):** Entrenado desde cero en 15M pares imagen-texto de literatura biomédica (PMC-15M dataset). Cubre múltiples modalidades: patología, radiología, microscopía. Zero-shot accuracy 58.3% en PathMNIST (pa-

tología) vs 42.1% de CLIP [50].

- **MedSigLIP (2024):** Adaptación de SigLIP específica para histopatología, pre-entrenada en corpus masivo de parches histológicos anotados. Arquitectura: ViT-Large/16 (304M parámetros), genera embeddings de 1152 dimensiones [38].

2.3.3 Métricas de Similitud y K-NN

Para embeddings normalizados ($\|\mathbf{v}\|_2 = 1$), las métricas dominantes son matemáticamente equivalentes:

$$\text{Cosine Similarity: } \cos(\theta) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|} = \mathbf{v}_1 \cdot \mathbf{v}_2 \quad (3.5)$$

$$\text{Inner Product: } \text{IP}(\mathbf{v}_1, \mathbf{v}_2) = \mathbf{v}_1 \cdot \mathbf{v}_2 \quad (3.6)$$

$$\text{Euclidean Distance: } \|\mathbf{v}_1 - \mathbf{v}_2\|_2^2 = 2 - 2(\mathbf{v}_1 \cdot \mathbf{v}_2) \quad (3.7)$$

Por lo tanto, maximizar producto interno \equiv maximizar similitud coseno \equiv minimizar distancia euclidiana. FAISS IndexFlatIP (Inner Product) es la elección estándar.

2.4 Large Language Models (LLMs) en Medicina

Los Large Language Models han demostrado capacidades emergentes en razonamiento complejo, comprensión de lenguaje natural y generación de texto coherente, lo que los posiciona como candidatos prometedores para asistencia clínica [42]. Esta sección examina su evolución arquitectural, adaptaciones al dominio médico, y los desafíos críticos de fiabilidad que motivan estrategias de mitigación como RAG.

2.4.1 Evolución

Era Pre-Transformer (2013-2017)

Los primeros intentos de NLP médico emplearon redes concurrentes (LSTM, GRU) entrenadas en corpus clínicos específicos. Sin embargo, su capacidad de generalización era limitada y el entrenamiento requería supervisión intensiva por tarea.

Transformers Bidireccionales: BERT (2018-2020)

BERT (Bidirectional Encoder Representations from Transformers) [7] revolucionó NLP mediante pre-entrenamiento no supervisado en masked language modeling. Su arquitectura encoder-only es óptima para tareas de comprensión (clasificación), pero limitada para generación.

Adaptaciones médicas tempranas:

- **BioBERT (KAIST, 2019):** Pre-entrenamiento adicional de BERT en PubMed (4.5M abstracts) + PMC (13.5M artículos full-text). Mejoras de 0.5-3% en NER biomédico y question-answering [22].
- **ClinicalBERT (MIT, 2019):** Especializado en notas clínicas de MIMIC-III. Superior en tareas sobre registros electrónicos de salud (EHR) [2].
- **PubMedBERT (Microsoft, 2020):** Entrenado desde cero (sin pesos de BERT general) exclusivamente en abstracts de PubMed, evitando conflictos de dominio [12].

Transformers Autoregresivos: Era GPT (2020-Presente)

GPT-3 (2020) demostró que modelos decoder-only entrenados en next-token prediction a escala masiva (175B parámetros, 300B tokens) exhiben capacidades de **in-context learning**: pueden resolver tareas nuevas mediante ejemplos en el prompt, sin fine-tuning [4]. Esta arquitectura domina generación de texto, diálogo y razonamiento multi-paso.

2.5 Sistemas de Apoyo Diagnóstico en Patología Mamaria

La aplicación de inteligencia artificial a la clasificación de cáncer de mama ha sido objeto de intensa investigación, particularmente desde la disponibilidad de datasets públicos curados y la maduración de arquitecturas de deep learning. Esta sección examina el estado del arte específico en el dominio de aplicación de este trabajo.

2.5.1 Clasificación Automática de Tejido Mamario

Vision Transformers en Histopatología

Desde 2021, Vision Transformers (ViT) han comenzado a superar CNNs en datasets grandes, capturando dependencias globales mediante mecanismos de atención:

- **Shao et al. (2021):** TransMIL, arquitectura transformer para Multiple Instance Learning en WSI. Alcanza 92.1% AUC en CAMELYON16, superando CNN-based MIL [40].
- **Wang et al. (2022):** CTransPath, pre-entrenado en 15M parches histopatológicos. Supera ResNet-50 en 7/9 datasets de clasificación tumoral [43].
- **Chen et al. (2024):** UNI foundation model (1B parámetros) logra state-of-the-art en múltiples benchmarks incluyendo ICIAR 2018 con 92.8% accuracy usando solo clasificador lineal [5].

Aunque ViTs y foundation models alcanzan accuracies altos, deben lidiar con la limitación de caja negra. Un accuracy de 92% significa 8% de errores sin mecanismo para que el patólogo identifique cuándo desconfiar del modelo.

Explicabilidad e Interpretabilidad en Patología

La interpretabilidad es reconocida como requisito crítico para adopción clínica [30]:

- **Grad-CAM y Variantes:** Mapas de activación que resaltan regiones importan-

tes para la predicción. Limitación: muestran "dónde" miró el modelo, no "por qué" decidió [39].

Las explicaciones visuales (heatmaps, prototipos) requieren que el patólogo interprete representaciones abstractas.

2.5.2 Integración de RAG y LLM en Aplicaciones Médicas

La combinación de recuperación de información con modelos de lenguaje generativos es emergente en medicina.

Trabajos Pioneros

- **MedRAG (Tang et al., 2024):** Sistema RAG textual para question-answering médico. Recupera pasajes de PubMed y alimenta GPT-4. Mejora accuracy de 61.3% (GPT-4 solo) a 68.9% (MedRAG) en MedQA [51].
- **BiomedGPT (Zhang et al., 2023):** Modelo multimodal que integra texto, imágenes radiológicas e histopatología. Pre-entrenado en 2.5M pares imagen-texto médicos. Zero-shot classification en PathMNIST: 62.1% [49].
- **Med-Flamingo (Moor et al., 2023):** Adaptación de Flamingo (DeepMind) para medicina, con capacidad de few-shot learning en tareas visuales. Alcanza 56.3% en VQA-RAD (visual question answering radiológico) con solo 4 ejemplos [29].
- **LLaVA-Med (Li et al., 2023):** Extensión de LLaVA (2.4.4) fine-tuneada en 60K conversaciones médicas multimodales sintéticas. Puede responder preguntas sobre radiografías y reportes simultáneamente [24].
- **PathChat (Sun et al., 2024):** LLM conversacional específico para patología, entrenado en 456K pares pregunta-respuesta extraídos de libros de texto. No implementa RAG; el conocimiento está en los pesos del modelo [25].

Todos operan en modalidad única (RAG textual O visual, no ambos simultáneamente). Ninguno específico para patología digital. La integración genuina de RAG visual y textual es prácticamente inexplorada.

Por lo tanto, no existen muchos sistemas publicados que combinen (1) RAG visual en histopatología con recuperación de parches similares, (2) RAG textual de guías clínicas, y (3) generación de respuestas en lenguaje natural que sintetice ambas fuentes.

2.6 Análisis Comparativo de Trabajos Relacionados

La Tabla 6.4 sintetiza las características de sistemas relevantes en comparación con esta propuesta:

Table 6.4: Comparación de sistemas de apoyo diagnóstico en patología mamaria.

Sistema	Año	Enfoque
Aresta et al.	2019	Ensemble CNN
Yan et al.	2020	DenseNet + Attention
CTransPath	2022	Vision Transformer pre-entrenado
UNI	2024	Foundation model
PathChat	2024	LLM conversacional
Este Trabajo	2025	RAG Visual + Textual + LLM

2.6.1 Gaps de Investigación

Falta de Interpretabilidad

Aunque técnicas como Grad-CAM proporcionan visualizaciones, no traducen features abstractas a terminología diagnóstica. Un mapa de activación que resalta una región no explica si el modelo detectó "atipia nuclear", "invasión estromal" o "artefacto de tinción". RAG visual aborda esto mostrando casos de referencia etiquetados que el patólogo puede interpretar directamente.

Ausencia de Actualización Incremental

El conocimiento médico evoluciona continuamente: NCCN actualiza guías trimestralmente, nuevos biomarcadores se descubren, casos extremadamente raros se publican. Fine-tuning de modelos grandes ($>1B$ parámetros) para cada actualización es insostenible. RAG permite agregar documentos o imágenes al índice sin modificar los pesos del modelo, con costo $O(1)$ por inserción.

Escasa Interactividad con Expertos

Los patólogos no son capaces de realizar correcciones a clasificaciones que el modelo realiza. Además, las soluciones antes vistas no integran razonamiento entre modelos de lenguaje y la evidencia obtenida al clasificar cada parche.

Limitada Integración RAG Visual + Textual

La medicina combina evidencia morfológica (imágenes) con conocimiento teórico (literatura). Un patólogo consultando un caso difícil revisa simultáneamente atlas histológicos (RAG visual) y guías de práctica clínica (RAG textual). Ningún sistema automatizado replica esta integración.

2.6.2 Posicionamiento de Este Trabajo

Este trabajo integra un RAG visual-textual multimodal específica para patología digital, con las siguientes contribuciones:

1. **RAG Visual en Histopatología Mamaria:** Recuperación basada en embeddings (MedSigLIP) para clasificación interpretable de tejido mamario en dataset ICIAR 2018. A diferencia de clasificadores end-to-end, cada predicción se acompaña de los k casos más similares del dataset de referencia, permitiendo validación visual directa por el patólogo.
2. **Arquitectura Dual RAG:** Integración de RAG visual (parches similares) y RAG textual (fragmentos de guías NCCN/WHO) en un pipeline unificado. El

LLM generador sintetiza ambas fuentes en respuestas coherentes que combinan evidencia morfológica y documental.

3. **Sistema Generador-Evaluador:** Implementación de arquitectura multi-agente de dos niveles (generador ligero → evaluador médico especializado) que equilibra latencia y calidad de respuesta. El supervisor es el encargado de verificar que la incorporación del modelo evaluador sea correcta.
4. **Human-in-the-Loop con Retroalimentación:** Flujo de trabajo donde correcciones del patólogo se almacenan y utilizan para actualización incremental del índice FAISS, creando un ciclo de mejora continua sin reentrenamiento completo del modelo.

Diseño e Implementación de la Solución

El desarrollo de herramientas de apoyo diagnóstico para patólogos requiere de un sistema capaz de (1) recuperar casos morfológicamente similares de una base de conocimiento validada por expertos, y (2) proporcionar asistencia conversacional contextualizada que responda consultas específicas en el dominio oncológico y de cáncer de mama. Este capítulo presenta el diseño conceptual de la solución propuesta, estructurada en cuatro módulos principales que operan de manera coordinada para asistir al patólogo en el análisis de Whole Slide Images (WSI).

3.1 Ambiente de Desarrollo e Infraestructura

El núcleo del sistema (backend) fue desarrollado utilizando el lenguaje **Python 3.12.7**, seleccionado por su amplio ecosistema de librerías para Inteligencia Artificial.

3.1.1 Dependencias de Software

En lugar de listar la totalidad del entorno virtual, la Tabla [1.1](#) resume las librerías fundamentales que constituyen la base lógica del sistema.

Table 1.1: Principales librerías y frameworks utilizados.

Categoría	Librería Principal	Función en el Proyecto
Deep Learning	<code>torch</code> (PyTorch) <code>accelerate</code>	Entrenamiento e inferencia de modelos. Optimización del uso de hardware (GPU).
Procesamiento de Imágenes	<code>opencv-python</code> TIAToolbox	Manipulación básica de imágenes. Manipulación de las WSI.
LLM & RAG	<code>transformers</code> <code>faiss-cpu</code> <code>mcp</code>	Acceso a modelos (Hugging Face). Base de datos vectorial. Protocolo de comunicación con modelos.
Backend	<code>fastapi</code> / <code>Flask</code>	API y comunicación con frontend.

3.1.2 Hardware de Procesamiento

Debido a la alta carga computacional requerida para el procesamiento de imágenes gigapíxel (WSI) y la inferencia de modelos LLM, se utilizó una estación de trabajo de alto rendimiento con las siguientes especificaciones técnicas:

- **CPU:** Intel® Core™ i9-9980XE (18 núcleos físicos, 36 hilos lógicos, frecuencia base de 3.00 GHz).
- **Memoria RAM:** 128 GB DDR4.
- **Aceleración Gráfica:** 4 × NVIDIA RTX Titan (24 GB VRAM cada una), totalizando 96 GB de memoria de video para paralelismo de datos.
- **Sistema Operativo:** Ubuntu 20.04.6 LTS

El equipo seleccionado corresponde al de mejor prestaciones dentro de nuestro laboratorio. Por lo mismo, ningún componente fue elegido específicamente para cumplir un rol dentro de nuestro sistema. Hay que destacar la existencia de 4 tarjetas de video, que permiten ejecutar diferentes modelos en simultáneo, la existencia de 128 GB de RAM que favorecen el proceso de entrenamiento en batches y una CPU de buenas prestaciones que asegura un desarrollo fluido.

A pesar de contar con 96 GB de memoria de video, para entrenamiento e inferencia de

modelos existe la posibilidad de que no sea suficiente. Por lo tanto, se espera ejecutar modelos en inferencia mayores a 27B de forma cuantizada.

3.2 Adquisición y Preprocesamiento de Datos

La estrategia de datos se dividió en tres flujos independientes: datos para el entrenamiento del clasificador visual (RAG de imágenes), corpus de conocimiento para el RAG de texto, y datos clínicos reales para la validación de la prueba de concepto.

Dataset para Recuperación de Imágenes (Image RAG)

Para garantizar una precisión base en el sistema de recuperación de imágenes similares, se requirió un conjunto de datos estandarizado de histopatología mamaria teñida con Hematoxilina y Eosina (H&E).

Se seleccionó el dataset del *ICIAR 2018 Grand Challenge on Breast Cancer Histology Images (BACH)* [3]. Este conjunto, licenciado bajo CC BY-NC-ND, proporciona una base de referencia validada por expertos, mitigando el ruido de etiquetado.

Características del Dataset BACH:

- **Volumen:** 400 imágenes microscópicas (parches).
- **Distribución de Clases:** Balanceado equitativamente (100 imágenes/clase) en cuatro categorías: *Normal*, *Benigno*, *InSitu*, *Invasivo*.
- **Especificaciones Técnicas:** Formato .tiff, resolución 2048×1536 px, escala de píxel $0.42 \mu\text{m}/\text{px}$.
- **Etiquetado:** *Image-wise* (una etiqueta única por imagen completa).

Dado que este dataset proviene de un desafío técnico competitivo (Grand Challenge), las imágenes presentan una calidad curada, libre de artefactos de escaneo significativos (e.g, desenfoque, manchas, plegamiento de tejido), lo que lo hace idóneo para poblar la base de conocimiento inicial de la base de datos vectorial.

Si bien el dataset contiene imágenes de buena calidad, contar con sólo 400 imágenes (100/clase) lo hace bastante pequeño. Sus clases identifican casos de forma general, pero el College American of Pathologist identifica más sub-clases dentro de cada una. Otro problema que surge al utilizar este dataset, es que proviene de una única fuente. Esto puede llevar a problemas de generalización, aunque que si el modelo generador de embeddings es robusto este problema puede ser menor.

Corpus para RAG de Texto (Base de Conocimiento)

Para dotar al modelo de lenguaje de contexto médico especializado, se construyó una base de conocimiento documental. Se priorizó literatura en idioma inglés para maximizar la compatibilidad y rendimiento de los modelos de *embeddings* y LLMs actuales. El corpus incluye:

1. Guías clínicas y estándares (NCCN Guidelines, European Society of Radiology).
2. Literatura de referencia patológica (*Rosen's Breast Pathology*, *WHO Classifications of Tumors*).
3. Material educativo de organizaciones reconocidas (American Cancer Society, Susan G. Komen).

Para ver en detalle todos los documentos recolectados para nuestro corpus, ver el apéndice A [5.5](#).

Datos de Validación (Casos Reales)

Para evaluar el desempeño de la Prueba de Concepto en un entorno realista, se estableció una colaboración con *Environ SPA*. Esta colaboración facilitó el acceso a 40 Whole Slide Images (WSI) anónimas de casos reales de cáncer de mama.

- **Formato:** .svs (Scanner Aperio).
- **Resolución Media:** 59759×46051 píxeles (aprox. 2.7 Gigapíxeles).
- **Magnificación:** 20x, tinción H&E.

Un desafío que presenta usar imágenes reales, es que provienen con variabilidad en la calidad del escaneo. Un ejemplo claro, es la existencia de manchas o sombras debidas a la manipulación del laboratorio de origen. Esto presenta la oportunidad de mejorar nuestra etapa de pre procesamiento y de hacerla robusta a datos de diferentes laboratorios

3.3 Visión General del Sistema

El sistema propuesto adopta una arquitectura modular compuesta por cuatro componentes principales, cuya interacción se ilustra en la figura 3.1 :

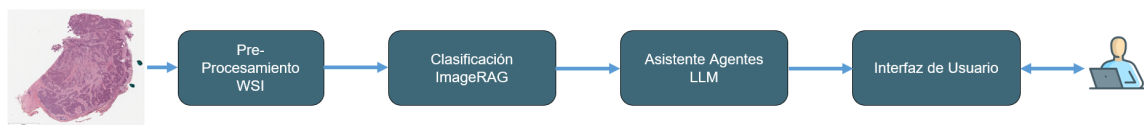


Figure 3.1: Arquitectura general del sistema de apoyo. El flujo comprende cuatro módulos principales: preprocesamiento de WSI, clasificación mediante ImageRAG, asistente conversacional LLM, e interfaz de interacción usuario-sistema con mecanismo de corrección

Componentes del Sistema

1. **Módulo Pre Procesamiento de WSI:** Componente encargado de la manipulación de la WSI, incluyendo segmentación de tejido, extracción de parches de tamaño fijo, y filtrado heurístico de regiones sin tejido (fondo, errores de escaneo, regiones fuera de foco).
2. **Módulo de Recuperación Visual (Image RAG):** Sistema de búsqueda por similitud basado en embeddings de parches de WSI. Utiliza un modelo de visión preentrenado en dominio médico (MedSigLIP) para codificar características morfológicas en un espacio latente denso, permitiendo recuperación eficiente de casos de referencia similares mediante búsqueda de vecinos más cercanos.

- 3. Módulo Conversacional (Agentes LLM y Modo Parche):** Asistente de lenguaje natural especializado en patología y cáncer de mama. Implementa una arquitectura de orquestación Generador-Evaluador, donde un modelo ligero genera respuestas preliminares que son posteriormente refinadas por un modelo especializado de mayor cantidad de parámetros. El sistema se refuerza mediante un RAG textual, accediendo a una base de conocimiento de guías clínicas y literatura especializada. Además de la posibilidad de conversar de la forma antes mencionada, el sistema incorpora el Modo Parche, que permite al usuario seleccionar un parche de la WSI y realizarle consultas específicas sobre el mismo.
- 4. Interfaz de Usuario:** Aplicación web que integra los módulos anteriores, permitiendo al patólogo (a) visualizar WSI con overlays de clasificación automática, (b) corregir predicciones erróneas mediante interacción directa, (c) realizar consultas textuales generales o específicas sobre regiones de interés, (d) consultar por parches similares en la base de datos vectorial y (e) recibir respuestas contextualizadas basadas en la evidencia visual y documental existente.

Flujo Operativo del Sistema

El flujo típico de uso del sistema comprende las siguientes etapas:

- 1. Carga de WSI:** El patólogo carga una WSI (típicamente archivo .svs de 2-3 gigapíxeles).
- 2. Preprocesamiento Automático:**
 - El sistema segmenta automáticamente regiones de tejido mediante el método Otsu .
 - Se extraen parches de 448×448 píxeles sin superposición.
 - Se aplican filtros heurísticos multi-criterio para descartar parches de baja calidad (fondo blanco, desenfoco, manchas).
- 3. Clasificación mediante RAG Visual:**

- Cada parche válido se codifica usando MedSigLIP, generando un embedding de 1152 dimensiones.
- El embedding se compara con el índice FAISS poblado con el dataset ICIAR 2018 BACH.
- Se recuperan los $k = 5$ vecinos más cercanos.
- La categoría se asigna mediante votación ponderada por similitud coseno.

4. Visualización y Corrección:

- La WSI clasificada se presenta en la interfaz web con overlays de colores según categoría.
- El patólogo puede seleccionar parches individuales para inspeccionar detalles.
- Si detecta errores de clasificación, puede reasignar manualmente la categoría correcta.
- Las correcciones se almacenan para futura mejora del modelo.

5. Consulta al Asistente LLM:

- El patólogo formula preguntas en lenguaje natural (generales o sobre parches específicos).
- El sistema recupera contexto relevante desde el RAG textual (guías clínicas), o del parche según sea el caso.
- Se incorporan estadísticas de la clasificación actual como contexto adicional.
- Un modelo ligero (generador) produce una respuesta preliminar.
- Un modelo especializado (evaluador) valida, corrige o amplía la respuesta.
- La respuesta final se presenta al usuario.

6. Iteración: El patólogo puede continuar navegando la WSI, corrigiendo clasifica-

ciones y formulando nuevas consultas, manteniendo un contexto conversacional mediante una memoria de historial.

3.4 Diseño del Módulo Pre Procesamiento de WSI

Este módulo procesa Whole Slide Images (WSI) mediante un flujo de trabajo multi-etapa que enfrenta el desafío de analizar imágenes de gigapíxeles manteniendo la relevancia diagnóstica y eficiencia computacional.

Inicialización y Lectura de WSI

La figura 4.2 muestra una vista general del pipeline:

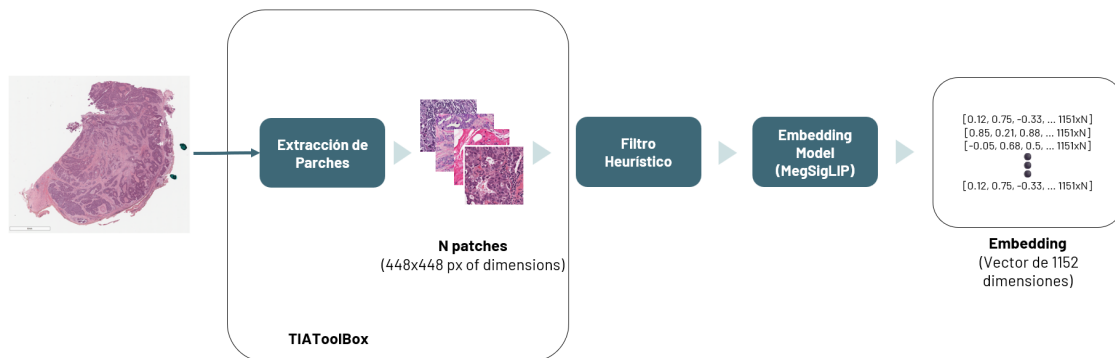


Figure 4.2: Arquitectura del pipeline de pre-procesamiento. El sub-módulo de TIAToolBox se encarga de leer la WSI, aplicar segmentación Otsu y recortar los parches, mientras que el filtrado heurístico y la generación de embeddings fueron incorporaciones externas.

Para el manejo de las WSI, se utilizó el framework **TIAToolbox** [32] para instanciar un objeto lector que maneja la estructura piramidal de los archivos .svs o .tiff, permitiendo un mapeo de memoria eficiente sin cargar la imagen completa en RAM. Si bien existen diferentes herramientas para manejar la WSI, TIAToolbox fue la herramienta que menor complejidad presentó, y se adecuó correctamente a nuestros requerimientos (como segmentación Otsu y manejo de coordenadas por parche).

Segmentación y Extracción de ROI

Para distinguir regiones de tejido biológico del fondo (vidrio, etiquetas, bordes), se implementa el método de **Otsu** para la binarización y separación automática tejido-fondo.

```
1 extractor = SlidingWindowPatchExtractor(  
    wsi_path,  
    input_mask="otsu",  
    min_mask_ratio=0.7  
)
```

Código 3.4.1 : Configuración del extractor de parches

El parámetro `min_mask_ratio=0.7` asegura que cada parche extraído contenga al menos un 70% de tejido biológico. Se utiliza una ventana deslizante (*sliding window*) para extraer parches de **448 × 448 píxeles** sin superposición, dimensión alineada con los requerimientos del modelo MedSigLIP.

Control de Calidad (Filtrado Heurístico)

Aunque la máscara Otsu descarta gran parte del fondo, pueden persistir falsos positivos: artefactos de escaneo, burbujas, plegamiento de tejido o zonas con tinción deficiente. Para asegurar una extracción de parches robusto, se implementó un filtro heurístico multi-criterio:

```
def is_blank_patch(  
    patch,  
    white_ratio_thresh=0.60,  
    black_ratio_thresh=0.05,  
5    sat_mean_thresh=25,  
    gstd_thresh=15,  
    structure_diff_thresh=8,  
    colorfulness_thresh=8,  
    mean_intensity_thresh=25,  
10 ):
```

Código 3.4.2 : Firma de la función de filtrado de parches

Justificación de los Umbrales:

- `white_ratio > 0.60`: Parches con más del 60% de píxeles blancos son probablemente fondo.
- `black_ratio >= 0.05`: Mayor al valor indicado quiere decir que el parche es negro.
- `sat_mean < 25`: Valores de saturación bajos indican tinción H&E deficiente o artefactos.
- `gstd_thresh = 15`: Desviación de estándar mínimo de escala de grises. Un valor menor indica parches sin estructuras celulares.
- `structure_diff = 8`: Diferencia mínima entre la imagen original y Gaussian-blurred. Valores bajos indican la ausencia de fillos y texturas.
- `colorfulness = 8`: Variación mínima de color en LAB (canales a^* y b^*). Las WSI deben tener colores rosa/morados distintivos.
- `mean_intensity = 25`: Media mínima de intensidad en escala de grises. Parches demasiados oscuros (< 25) pueden ser errores de procedimiento de digitalización.

3.5 Diseño del Módulo de Recuperación Visual (Image RAG)

Para el problema de recuperación de casos similares en parches, se evaluaron tres enfoques principales, cuyos trade-offs se resumen en la tabla 5.2

Table 5.2: Comparación de enfoques para clasificación de imágenes.

Enfoque	Ventajas	Desventajas	Decisión
Clasificación directa CNN	Arquitectura simple end-to-end.	Sin interpretabilidad. Requiere reentrenamiento completo.	Descartado
Fine-tuning de modelo fundacional	Alta precisión con conocimiento preentrenado.	Costoso en cómputo. Necesita datasets grandes (>10K imágenes).	Descartado
RAG con embeddings preentrenados	Interpretable y escalable. Eficiente con modelo preentrenado (MedSigLIP).	Depende de calidad del dataset. Búsqueda puede ser lenta.	Incorporado

La selección del RAG visual se fundamenta en tres criterios críticos para el contexto clínico:

1. **Interpretabilidad Clínica:** Un sistema de apoyo diagnóstico debe permitir al patólogo validar las sugerencias del modelo mediante evidencia visual. RAG retorna explícitamente los k casos más similares del dataset de referencia, permitiendo comparación morfológica directa. Esto es imposible con clasificadores tipo "caja negra" donde la decisión se basa en pesos neuronales ininterpretables.

2. **Escalabilidad sin Reentrenamiento:** La incorporación de nuevos casos diagnósticos (incluyendo correcciones del patólogo) debe ser posible de manera incremental. RAG permite agregar nuevos embeddings al índice FAISS sin necesidad de reentrenar el modelo completo, un proceso que en fine-tuning requeriría horas o días de cómputo GPU.
3. **Eficiencia Computacional para Prototipo:** El fine-tuning de modelos Vision Transformer grandes (e.g., ViT-L/16 con 300M+ parámetros) requiere recursos computacionales y datasets que exceden el alcance de esta prueba de concepto. El uso de embeddings preentrenados de MedSigLIP—un modelo ya optimizado en millones de pares imagen-texto médicos—reduce drásticamente el costo de desarrollo sin sacrificar precisión diagnóstica.

Para la implementación del clasificador ImageRAG, fue de mucha utilidad (e inspiración) el trabajo realizado por **Konstantinos et al.** [37]

3.5.1 Arquitectura del Sistema RAG Visual

El módulo de recuperación visual opera en dos fases: **construcción offline del índice y búsqueda de similitud en inferencia.**

3.5.1.1 Fase 1: Construcción del Índice Vectorial

La población inicial de la base de datos vectorial se realiza procesando el dataset de referencia ICIAR 2018 BACH mediante el pipeline ilustrado en la figura 5.3:

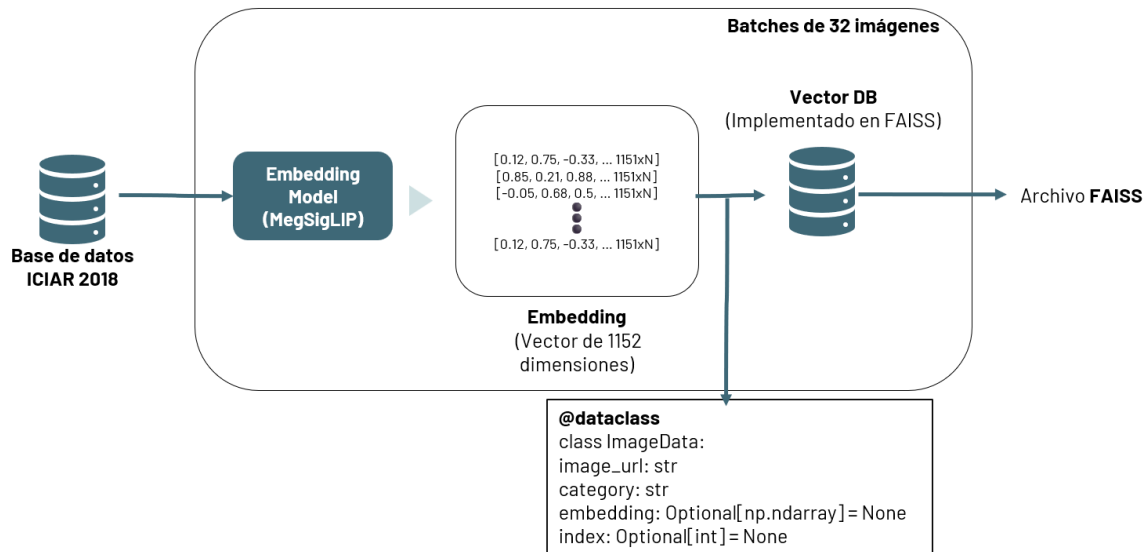


Figure 5.3: Población inicial del ImageRAG con dataset ICIAR 2018. En batches de 32 imágenes, se calcula su embedding, almacenan los datos de ubicación y categoría de cada imagen, y finalmente se insertan en el índice FAISS.

Etapas del Pipeline de Construcción de RAG:

1. **Carga del Dataset ICIAR 2018:** Se cargan las 400 imágenes del dataset (100 por clase: Normal, Benign, InSitu, Invasive), cada una de resolución 2048×1536 píxeles a $0.42 \mu\text{m}/\text{px}$.
2. **Preprocesamiento de Imágenes:** Cada imagen se redimensiona a 448×448 píxeles (resolución de entrada de MedSigLIP).
3. **Generación de Embeddings con MedSigLIP:** Se seleccionó MedSigLIP [38] como encoder visual por tres razones:
 - **Especialización en Dominio Médico:** Pre-entrenado en millones de pares imagen-texto de literatura biomédica (PubMed, libros de texto médicos), a diferencia de modelos genéricos entrenados solo en ImageNet.
 - **Arquitectura Optimizada:** Basado en CLIP con modificaciones para imágenes médicas, generando embeddings de 1152 dimensiones que capturan

tanto características morfológicas como semánticas.

- **Desempeño Validado:** Reporta métricas superiores en benchmarks de patología digital comparado con ResNet-50, ViT base, y otros encoders no especializados [38].

Cada imagen \mathbf{I} se transforma en un vector denso:

$$\mathbf{v} = \text{MedSigLIP}(\mathbf{I}) \in \mathbb{R}^{1152} \quad (5.1)$$

4. **Normalización L2:** Los embeddings se normalizan para facilitar comparación mediante similitud coseno:

$$\mathbf{v}_{\text{norm}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}, \quad \|\mathbf{v}_{\text{norm}}\|_2 = 1 \quad (5.2)$$

Ventajas de la Normalización L2:

- Transforma el producto interno en similitud coseno: $\langle \mathbf{u}, \mathbf{v} \rangle = \cos(\theta)$ donde θ es el ángulo entre vectores.
 - Mejora la estabilidad numérica de la búsqueda FAISS.
 - Facilita la agregación de múltiples vecinos mediante suma ponderada.
5. **Indexación en FAISS:** Los vectores normalizados se insertan en un índice IndexFlatIP (Inner Product sobre vectores normalizados):

- **Tipo de Índice:** IndexFlatIP implementa búsqueda exhaustiva con complejidad $O(nd)$ por consulta, donde $n = 400$ (tamaño del dataset) y $d = 1152$ (dimensionalidad).
- **Justificación:** Aunque existen variantes aproximadas más rápidas (e.g., IndexIVFFlat con cuantización), se priorizó **precisión máxima** sobre velocidad dada la criticidad diagnóstica. Para $n = 400$, la búsqueda exhaustiva

toma <10ms en CPU moderna.

- **Escalabilidad Futura:** Si el dataset crece a decenas de miles de imágenes, será necesario migrar a índices aproximados como HNSW o IVF con cuantización PQ.

6. **Procesamiento por Lotes (Batch Processing) e Indexado** Para optimizar la memoria y la eficiencia computacional, se emplea una estrategia secuencial por lotes con un *batch size* de 32 imágenes. Para cada lote, el sistema realiza:

- (a) Carga y preprocesamiento de la imagen acorde a MedSigLIP.
- (b) Generación de embeddings mediante el codificador.
- (c) Acumulación de embeddings y metadatos (etiquetas del dataset ICIAR 2018).
- (d) Ejecución de una inserción masiva (*bulk insertion*) en el índice FAISS.

7. **Almacenamiento de Metadatos:** Paralelamente al índice FAISS, se mantiene un archivo .pkl que mapea cada posición del índice a sus metadatos clínicos:

```
@dataclass
class ImageData:
    image_url: str
    category: str
    embedding: Optional[np.ndarray] = None
    index: Optional[int] = None
```

Código 3.5.3 : Estructura de metadatos por imagen.

Los embeddings se almacenan en FAISS (optimizado), la metadata en cambio, almacena referencias y campos adicionales no numéricos. Esta decisión favorece a la velocidad de recuperación.

Este mapeo es fundamental para: (a) recuperar la etiqueta diagnóstica de los vecinos encontrados, (b) cargar las imágenes originales para visualización, y (c)

actualizar el índice con nuevos casos sin perder trazabilidad.

8. **Persistencia Dual** Para almacenar la base vectorial, junto a los datos asociados a cada imagen, se implementa una estrategia de persistencia dual para una carga parcial eficiente.

Archivos Generados:

- `filepath.faiss`: Estructura serializada del índice con vectores optimizados.
- `filepath.pkl`: Metadatos serializados que mapean cada posición del índice a su información clínica.

3.5.1.2 Fase 2: Búsqueda por Similitud (Inference)

La figura 5.4 representa el procedimiento de recuperación y clasificación de ImageRAG.

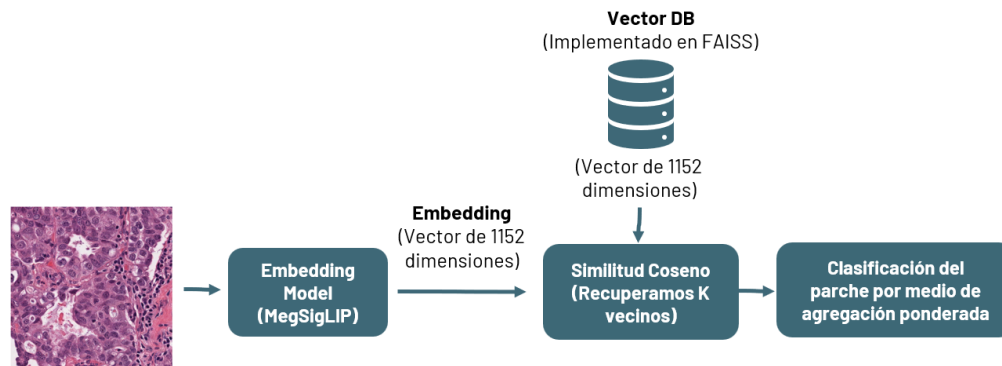


Figure 5.4: Recuperación y clasificación de parches con ImageRAG. Dado un parche, obtenemos su embedding con MedSigLIP y a través de agregación ponderada obtenemos clasificación (que corresponde a la categoría con mayor puntaje).

El módulo RAG ejecuta los siguientes pasos para lograr una eficiente recuperación y clasificación:

1. **Extracción de Parches:** La WSI se segmenta en parches de 448×448 píxeles **sin superposición** (stride = patch size). Esta dimensión se seleccionó por dos razones:
 - **Compatibilidad con MedSigLIP:** El modelo fue entrenado y optimizado con imágenes de entrada 448×448 px, demostrando el mejor desempeño en tareas médicas a esta resolución.
2. **Encoding:** Cada parche válido $\mathbf{P}_{\text{query}}$ se procesa con el mismo encoder MedSigLIP utilizado en la fase de indexación, generando su embedding normalizado $\mathbf{q} \in \mathbb{R}^{1152}$.
3. **Búsqueda K-NN en FAISS:** Para cada embedding de consulta \mathbf{q} , se recuperan los $k = 5$ vecinos más cercanos del índice mediante:

$$\mathcal{N}_k(\mathbf{q}) = \underset{v_i \in \mathcal{D}}{\text{arg top-}k} \langle \mathbf{q}, \mathbf{v}_i \rangle \quad (5.3)$$

donde \mathcal{D} es el conjunto de embeddings del dataset ICIAR y $\langle \cdot, \cdot \rangle$ denota el producto interno (equivalente a la similitud coseno para vectores normalizados).

Justificación de $k = 5$

- **Robustez Estadística:** Cinco vecinos proporcionan suficiente evidencia para votación ponderada sin introducir ruido de casos distantes.
 - **Eficiencia:** Valores mayores de k (probados: 10, 15) no mejoraron la precisión de clasificación en experimentos preliminares.
4. **Clasificación por Agregación Ponderada:** La clasificación final utiliza una estimación de máximo a posteriori basada en votación ponderada por similitud:

$$\text{score}(c) = \sum_{i:\text{label}_i=c} \frac{\text{sim}_i}{\sum_i \text{sim}_i} \quad (5.4)$$

La predicción final es:

$$\hat{c} = \arg \max_{c \in \{\text{Normal, Benign, InSitu, Invasive}\}} \text{score}(c) \quad (5.5)$$

```
def get_category_suggestions(...)  
    ...  
    ...  
4     for result in similar_images:  
        category = result['category']  
        weight = result['similarity_score']  
  
        if category not in category_scores:  
9             category_scores[category] = 0  
            category_scores[category] += weight  
            total_weight += weight  
  
        if total_weight > 0:  
14            for category in category_scores:  
                category_scores[category] /= total_weight  
  
        ...  
    }
```

Código 3.5.4 : Extracto de clasificación por suma ponderada de vecinos. Se aprecia el algoritmo utilizado para obtener el puntaje.

El score normalizado se interpreta como **confianza de la predicción**, proporcionando al patólogo una medida cuantitativa de la certeza del modelo.

5. **Presentación de resultados:** Los resultados de todos los parches clasificados se presentan en un archivo JSON:

```
"patches": [  
  2 {  
    "index": 24,  
    "category": "invasive",  
    "x": 27328,  
    "y": 3136,  
    7 "width": 448,  
    "height": 448,  
    "confidence": 0.601  
  },  
  ...  
  12 ],  
  "stats": {  
    "normal": 475,  
    "benign": 295,  
    "invasive": 1060,  
    17 "insitu": 91  
  },  
  "image_dimensions": {  
    "width": 47807,  
    "height": 47634  
    22 },  
  "total_patches": 2168
```

Código 3.5.5 : JSON con información de una WSI después del proceso de clasificación.

Tras clasificar todos los parches, se computan estadísticas a nivel de WSI completa, las cuales en un futuro serán parte del contexto entregado a los modelos LLM. Este formato permite: (a) renderizado eficiente en la interfaz web, (b) acceso a la información de forma eficaz, y (c) almacenamiento persistente sin reprocesar la WSI.

3.5.2 Manejo de Desbalanceo de Clases

El dataset ICIAR 2018 presenta una distribución balanceada (100 imágenes por clase), lo cual mitiga problemas de sesgo en la recuperación. Sin embargo, se identificaron dos desafíos:

- **Similitud Morfológica entre Clases:** Las clases *Normal* y *Benign* comparten características visuales [33], lo que puede causar confusión en casos limítrofes. Para efectos de visualización en la interfaz, ambas clases se representan con el mismo código de color (gris), dado que ninguna requiere intervención terapéutica inmediata.
- **Representatividad Limitada:** El sistema solo puede clasificar en las cuatro categorías del dataset ICIAR. Casos que presentan características fuera de estas categorías (e.g., lesiones papilares, carcinoma mucinoso) no serán correctamente identificadas.

3.5.3 Limitaciones del Diseño RAG

- **Dependencia de la Calidad del Dataset:** El sistema está intrínsecamente limitado por la calidad y cobertura del dataset ICIAR. Con solo 400 imágenes, la variabilidad morfológica es limitada comparada con la práctica clínica real donde existen sub-clasificaciones más finas. Por ejemplo, el American College of Pathologist reconoce más de 20 subtipos de carcinoma mamario, mientras que ICIAR solo cubre 4 categorías amplias.

- Complejidad Computacional de Búsqueda Exhaustiva:** El índice IndexFlatIP tiene complejidad temporal $O(n \cdot d)$ por consulta, donde n es el número de vectores indexados y $d = 1152$ la dimensionalidad. Si se expande el dataset a 100.000 imágenes, la búsqueda exhaustiva tomaría ~ 1 -2 segundos por parche, haciendo el sistema no apto para WSI con miles de parches.

3.6 Diseño del Módulo Conversacional (Agentes LLM)

La generación de respuestas clínicamente precisas en patología presenta desafíos únicos: (1) el modelo debe dominar terminología médica especializada, (2) debe evitar "hallucinations" que podrían causar decisiones diagnósticas erróneas, y (3) debe mantener consistencia con guías clínicas validadas. Para abordar estos requisitos, se diseñó una arquitectura de orquestación basada en el patrón **Generador-Evaluador**. La figura 6.5 resume la arquitectura propuesta.

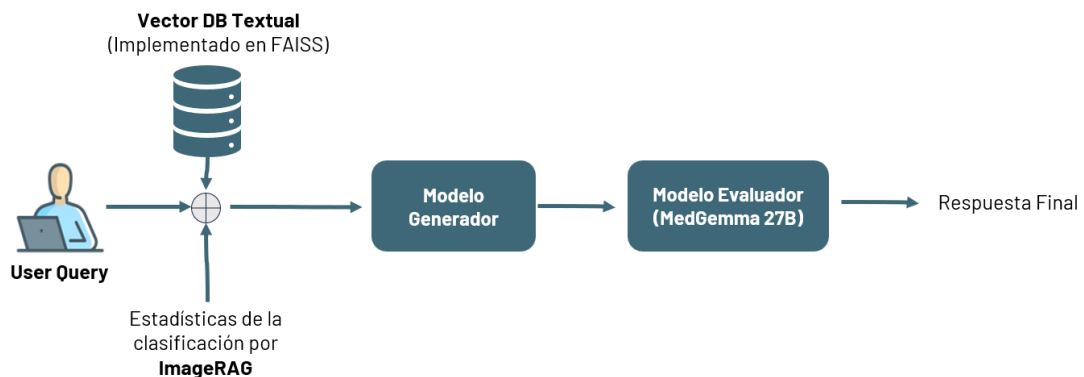


Figure 6.5: Arquitectura del sistema de chat con evaluación en cascada. El flujo incluye recuperación de contexto (RAG textual + estadísticas WSI), generación inicial por modelo ligero y evaluación por modelo especializado (modelo que reflexiona a través del modo **Thinking**).

3.6.1 Motivación del Diseño Multi-Agentes

En lugar de utilizar un único modelo LLM de gran escala (e.g, GPT-4, Claude 3.5), se

optó por una arquitectura de dos etapas por las siguientes razones:

1. **Latencia y Costo:** Modelos comerciales grandes tienen latencias de 3-10 segundos por respuesta y costos de \$0.01-0.03 por consulta. Para un patólogo que analiza 20-30 casos diarios, esto implica costos recurrentes significativos y tiempos de espera que interrumpen el flujo de trabajo.
2. **Privacidad de Datos:** El uso de APIs comerciales implica enviar datos potencialmente sensibles (detalles de casos, consultas diagnósticas) a servidores externos, violando regulaciones de privacidad médica. Un sistema local con Small Language Models (SLMs) mantiene todos los datos dentro de la infraestructura del hospital.
3. **Validación de Precisión Médica:** Los Small Languages Models (4-8B de parámetros) son rápidos y eficientes, pero pueden generar respuestas incompletas o inexactas en consultas complejas. Un modelo especializado más grande (27B parámetros en nuestro caso) actúa como supervisor de calidad, corrigiendo errores sin el costo de inferencia inicial.
4. **Objetivo Experimental:** Uno de los objetivos específicos de este trabajo es evaluar si los Small Languages Models pueden ser útiles en aplicaciones médicas reales cuando son reforzados con un evaluador especializado, versus requerir modelos comerciales masivos. Esta arquitectura permite cuantificar el valor agregado del evaluador en el capítulo de Resultados.

3.6.2 Componentes de la Arquitectura

El módulo conversacional comprende dos agentes con roles especializados:

3.6.2.1 Agente Generador (Small Language Model):

Este modelo recibe el contexto inicial, que incluye fragmentos recuperados del RAG de texto, estadísticas globales de la WSI clasificada y el historial de conversación (memo-

ria). Genera una respuesta preliminar rápida.

Para este trabajo se seleccionaron tres modelos ligeros (entre 4B-8B de parámetros) como generador inicial. Los criterios de selección fueron:

- **Especialización Médica:** Preferencia por modelos fine-tuneados en literatura biomédica.
- **Capacidad de Instrucción:** Modelo del tipo "instruct" capaz de seguir prompts estructurados y mantener formato de respuesta.

Los modelos seleccionados corresponden a MedGemma 4B, Meerkat 7B y OpenBioLLM 8B ([8, 20, 38]). Ver 4.1.2 para más detalles.

Contexto Proporcionado al Generador:

El prompt enviado al modelo generador se enriquece con tres fuentes de contexto:

1. **RAG Textual:** Fragmentos recuperados de la base de conocimiento documental (que corresponde a otro de los objetivos específicos de este trabajo) mediante búsqueda semántica por similitud con la consulta del usuario.
2. **Estadísticas de la WSI actual:** Distribución de categorías clasificadas (e.g., "43 parches Invasive, 183 InSitu, 412 Benign, 1205 Normal"), proporcionado contexto cuantitativo sobre el caso en análisis.
3. **Historial de Conversación:** Últimas $N = 10$ respuestas del modelo para mantener coherencia contextual en diálogos multi-turno. Esta característica puede ser habilitada por el usuario.

3.6.2.2 Agente Evaluador (MedGemma 27B):

Modelo especializado de 27B de parámetros (cuantizado a INT4 para reducir VRAM) que actúa como supervisor de calidad médica. Recibe la respuesta del modelo 4B y utiliza un mecanismo de `is_thinking` para razonar sobre la validez médica de la respuesta antes de emitir el output final al usuario.

El evaluador recibe:

- La respuesta generada por el SLM.
- La pregunta original del patólogo.

Mecanismo de Razonamiento Explícito:

El evaluador utiliza un token especial <thinking> para realizar razonamiento intermedio antes de emitir la respuesta final.

Este mecanismo permite al evaluador:

- Detectar errores en la respuesta del generador
- Identificar omisiones de información crítica
- Ampliar detalles técnicos cuando sea necesario

3.6.3 Estrategia de Prompting

La interacción con los modelos se estructura mediante *prompts* especializados que condicionan el comportamiento y formato de respuesta del modelo. A continuación se presentan las plantillas utilizadas:

3.6.3.1 Prompt del Generador

```
Last 10 Chats (may or may not be useful):  
2 {history_text}  
  
WSI Analysis Context:  
- WSI: {request.wsi_name}  
- Statistics: {stats_summary}  
7  
Additional Context (RAG, may or may not be useful):  
- {ragInformation}  
  
Instructions:  
12 - Use the RAG information only if it provides clarity or relevant data  
.  
- Use the Last 10 Chats information only if it provides clarity or  
relevant data.  
- If the RAG is not useful, ignore it completely  
  
User Question:  
17 {request.message}
```

Código 3.6.6 : Estructura del prompt para el SLM generador

3.6.3.2 Prompt del Evaluador

```
You are an expert medical assistant. Your task is to provide accurate,  
complete, and clear medical answers.  
You will receive a user's question and a draft response.  
3 Your job is to provide the best possible final answer to the user's  
question, incorporating any useful information from the draft.  
Do not critique or comment on the draft - simply provide your best  
answer directly.  
  
User's question:  
  
8 {input}  
  
Draft response:  
  
{initial_response}  
13  
  
Please provide the best possible answer to the user's question.
```

Código 3.6.7 : Estructura del prompt para MedGemma 27B

3.6.4 Enriquecimiento Contextual mediante RAG Textual

Para mitigar el fenómeno de alucinaciones y fundamentar las respuestas del sistema en evidencia documental verificable, se implementó un sistema de Retrieval-Augmented Generation (RAG) textual que integra recuperación semántica con reordenamiento mediante cross-encoders.

3.6.4.1 Construcción de la Base de Conocimientos

El pipeline de construcción de la base vectorial consta de las siguientes etapas:

1. **Corpus Documental:** Se priorizó literatura médica en inglés de alta calidad y validez científica, incluyendo guías clínicas NCCN (National Comprehensive Can-

cer Network), clasificaciones WHO (World Health Organization), y textbooks de referencia en patología mamaria como Rosen's Breast Pathology. Se excluyeron fuentes no validadas (foros, blogs médicos, sitios de divulgación) para garantizar la fiabilidad de la información recuperada. El sistema implementa carga automatizada de documentos PDF mediante PyPDFLoader del framework LangChain.

2. **Segmentación Semántica (Chunking):** La fragmentación de documentos se realiza mediante `RecursiveCharacterTextSplitter` de LangChain, que implementa una estrategia jerárquica de separación basada en delimitadores naturales del lenguaje. Los parámetros de segmentación fueron optimizados considerando el balance entre contexto semántico y precisión de recuperación:

- **Tamaño de chunk:** 400 tokens (medidos con el tokenizador de `sentence-transformers/all-mpnet-base-v2`)
- **Overlap:** 80 tokens (20% de solapamiento para preservar continuidad contextual)
- **Separadores jerárquicos:** Dobles saltos de línea (`\n\n`), saltos simples (`\n`), puntos seguidos de espacio, puntos y comas, dos puntos, comas y espacios individuales, aplicados recursivamente en orden de prioridad

Esta estrategia de separación preserva la integridad de párrafos y oraciones, minimizando la fragmentación de conceptos médicos complejos que requieren múltiples oraciones para su comprensión.

3. **Generación de Embeddings Densos:** Cada chunk resultante se codifica mediante el modelo `sentence-transformers/all-mpnet-base-v2`, generando representaciones vectoriales densas de 768 dimensiones en espacio semántico continuo. Este modelo fue seleccionado por su desempeño superior en benchmarks de recuperación semántica (MS MARCO, Natural Questions) y su capacidad de capturar relaciones semánticas complejas en texto médico técnico. El proceso de embedding se optimiza mediante procesamiento por lotes (batch size = 8) con

aceleración GPU cuando está disponible.

4. **Indexación Vectorial:** Los embeddings generados se almacenan en una base de datos vectorial FAISS (Facebook AI Similarity Search), optimizada para búsqueda de vecinos más cercanos mediante algoritmos aproximados. Cada vector mantiene metadatos estructurados que incluyen nombre del archivo fuente (`filename`), número de página (`page`) y ruta del documento (`source`), permitiendo trazabilidad completa de la información recuperada. El índice FAISS se persiste en disco para reutilización en inferencia sin reentrenamiento.

La figura 6.6 representa el proceso de población inicial:

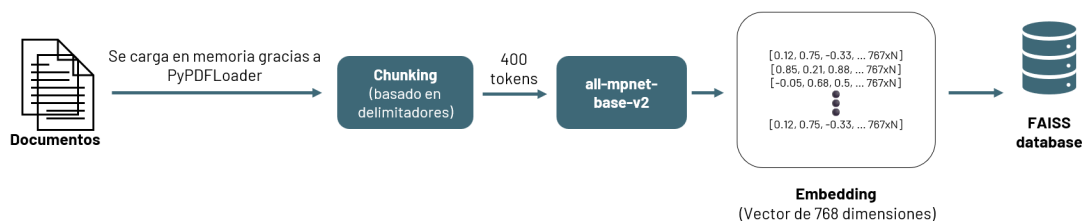


Figure 6.6: Esquema de población de RAG textual.

Recuperación en Tiempo de Consulta

El proceso de recuperación implementa una arquitectura de dos etapas para maximizar tanto recall como precisión:

1. **Embedding de la Consulta:** La pregunta del patólogo se codifica utilizando el mismo modelo de embeddings (`sentence-transformers/all-mpnet-base-v2`) para garantizar consistencia en el espacio vectorial compartido.
2. **Recuperación Inicial por Similitud:** Se ejecuta búsqueda de similitud coseno en el índice FAISS para recuperar los top- k candidatos iniciales (típicamente $k = 20$), privilegiando recall en esta etapa preliminar.

3. **Reordenamiento con Cross-Encoder (Reranking):** Los candidatos recuperados se reordenan mediante un modelo cross-encoder (`ms-marco-MiniLM-L-12-v2`) que evalúa pares query-documento de forma conjunta, generando scores de relevancia más precisos que la similitud coseno. Este modelo procesa la concatenación [query, documento] mediante transformers bidireccionales, capturando interacciones semánticas complejas entre la pregunta y el contenido. Del conjunto reordenado se seleccionan los top- k definitivos (típicamente $k = 3 - 5$) con mayor score de relevancia.
4. **Inyección Contextual:** Los chunks finales se formatean estructuradamente incluyendo metadatos de fuente, página y score de relevancia, e se insertan en el prompt del modelo generativo como contexto documental, permitiendo al LLM fundamentar sus respuestas en evidencia específica.

La figura 6.7 representa el proceso de recuperación de documentos:

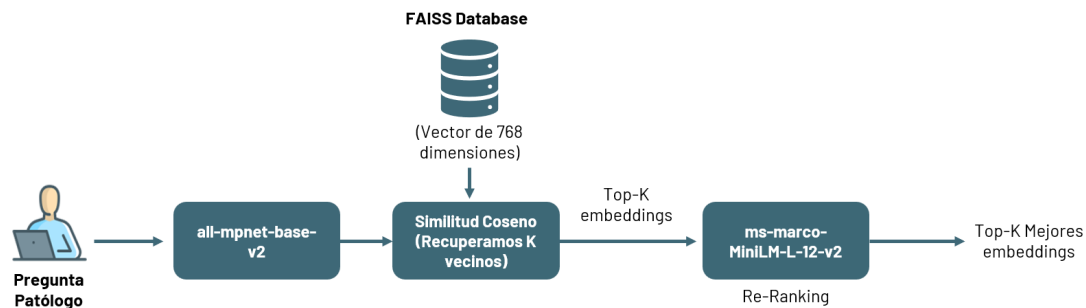


Figure 6.7: Recuperación de documentos desde base de datos vectorial. Los Top-K Mejores embeddings se concatenan al prompt del modelo generador, proporcionándole el contexto adecuado.

Esta arquitectura híbrida combina la eficiencia de la búsqueda vectorial densa (bi-encoder) con la precisión del reordenamiento semántico profundo (cross-encoder), resultando en recuperación de alta calidad para consultas médicas especializadas.

3.6.4.2 Manejo de Historial

Para mantener coherencia en diálogos multi-turno, se implementó un buffer circular de conversación:

- **Estructura de Datos:** Cola (queue) de tamaño máximo $N = 10$ mensajes.
- **Política FIFO:** Cuando se excede el límite, se descarta el mensaje más antiguo.

3.6.5 Consultas Multimodales sobre Parches Específicos

Además de consultas textuales generales, el sistema permite al patólogo seleccionar un parche específico de la WSI y formular preguntas sobre esa región concreta (e.g., «Qué tipo de células componen esta estructura?», «Por qué fue clasificado como Invasive este parche?»). El modelo encargado de analizar estas preguntas fue el «MedGemma 4B Multimodal», que destaca por:

- **Pocos parámetros:** Al tener pocos parámetros, ayuda a disminuir el costo computacional de ejecutar los modelos en el pipeline.
- **Entrenamiento en Contexto Médico:** El modelo de MedGemma cuenta con un sólido entrenamiento en imágenes-texto de medicina. Esta característica mejora la precisión de las respuestas.

Flujo de Consulta Multimodal:

1. El patólogo hace clic en un parche en la interfaz web
2. El frontend envía las coordenadas (x, y) del parche al backend
3. El sistema extrae la región de la WSI a máxima resolución
4. La imagen del parche + pregunta del usuario se envían al modelo MedGemma
5. El modelo genera la respuesta analizando visualmente la morfología del parche

3.6.6 Limitaciones del Diseño de Agentes LLM

- **Hallucinations:** A pesar del RAG textual y el evaluador, los LLMs pueden generar información falsa que suena plausible (especialmente en preguntas fuera del alcance del corpus).
- **Limitación de Idioma:** Los modelos están mayormente entrenados para inglés. Respuestas en español pueden presentar:
 - Menor fluidez gramatical
 - Respuestas terminadas abruptamente
 - Falta de precisión

Por lo tanto, toda interacción con el modelo se realiza en inglés.

- **Costo Computacional de Arquitectura Multi-Agente:** Ejecutar secuencialmente generador + evaluador aumenta significativamente la latencia. Por lo tanto, es crítico determinar si el modelo evaluador aporte a la respuesta final.
- **Mala elección del Prompt:** Un apartado poco estudiado, pero que en la actualidad está recibiendo mayor atención, trata sobre cómo afecta el prompt al comportamiento del modelo. Un prompt bien estructurado es capaz de guiar al modelo en su respuesta, y por lo tanto, obtener resultados con mayor precisión. Ya que este trabajo no tiene como finalidad estudiar los prompts más efectivos para el análisis de WSI, existe la posibilidad de que los prompts escogidos no sean adecuados en el contexto médico.
- **Ruido en RAG Textual:** El RAG textual tiene como objetivo facilitar un contexto sólido y actual al modelo LLM sobre las preguntas en cuestión. Sin embargo, si nuestro corpus textual carece de información relevante para la pregunta, entregará información que no hará más que confundir al modelo LLM.

3.7 Interfaz de Usuario

La interfaz de usuario (UI) tiene el rol de ser la capa de interacción entre el backend de Deep Learning y el patólogo. La filosofía de diseño prioriza la **interpretación visual**, asegurando que las predicciones del modelo-específicamente la clasificación de los parches de la Whole Slide Image (WSI)-se representen en un espacio coherente y fácilmente navegable. La arquitectura del frontend se construyó como una Single Page Application (SPA) usando tecnologías web estándares como HTML5, CSS3 y ES6+ Javascript, comunicándose con el backend a través de una RESTful API.

Visualización Piramidal De Imagen

Para manejar la escala de gigapíxeles inherente en las imágenes hispatológicas, el sistema integra **OpenSeaDragon**, una librería open-source de JavaScript especializada en el renderizado de Deep Zoom Images (DZI).

En vez de cargar la WSI completa en memoria, lo cual excedería los límites de los recursos del navegador, el visualizador consume la imagen en una estructura piramidal. La implementación inicializa la instancia del visualizador, que captura de forma asíncrona 'láminas' basadas en el nivel de zoom que tenga el usuario y de las coordenadas del viewport. Esto asegura una navegación de baja latencia incluso para imágenes de 100.000×100.000 píxeles.

Sistema De Overlay Para Los Parches

Las inferencias del modelo clasificador RAG son proyectadas en la WSI usando un enfoque **DOM-overlay**. Una vez inicializada la página web, el cliente captura la información de la clasificación de la WSI desde el endpoint `/wsi/wsi_name`. Esta data se renderiza como una capa de elementos interactivos sobrepuesto sobre el canvas de OpenSeaDragon.

La lógica de visualización sigue un código de color específico para representar las categorías patológicas:

- **Invasive:** Rojo (#FF0000)
- **In Situ:** Amarillo (#FFFF00)
- **Benign/Normal:** Plomo (#808080)

Cada overlay mantiene un mapeo espacial $(x, y, width, height)$ relativo a las coordenadas de la imagen WSI. Para optimizar la experiencia del usuario, se implementó una estrategia de delegación de eventos para manejar la interacción del usuario con los parches, gatillando un 'panel de inspección' para mostrar la metadata del parche sin obstaculizar la región de interés (ROI).

3.7.1 Flujo De Trabajo Human-In-The-Loop

Reconociendo la necesidad de los expertos de corregir posibles fallos en el proceso de clasificación de los parches, la interfaz presenta un mecanismo de Human-In-The-Loop. Esta característica permite a los patólogos intervenir en el resultado mostrado, interviniendo además en el algoritmo de clasificación.

Cuando un parche en específico es seleccionado, el usuario puede acceder a un apartado de modificación (Ver figura 7.9). El frontend captura la corrección del patólogo y la transmite a la API a través de un requerimiento PUT, actualizando la data de clasificación del parche, además, el sistema se encarga de extraer el parche de la WSI y almacena en un archivo csv su nombre y nueva categoría, para utilizarlo en una posterior mejora del Image RAG. Esta funcionalidad transforma el software de un visualizador pasivo a una herramienta interactiva de anotación y verificación, facilitando la mejora del modelo. La figura 7.8 presenta un esquema de lo mencionado anteriormente:

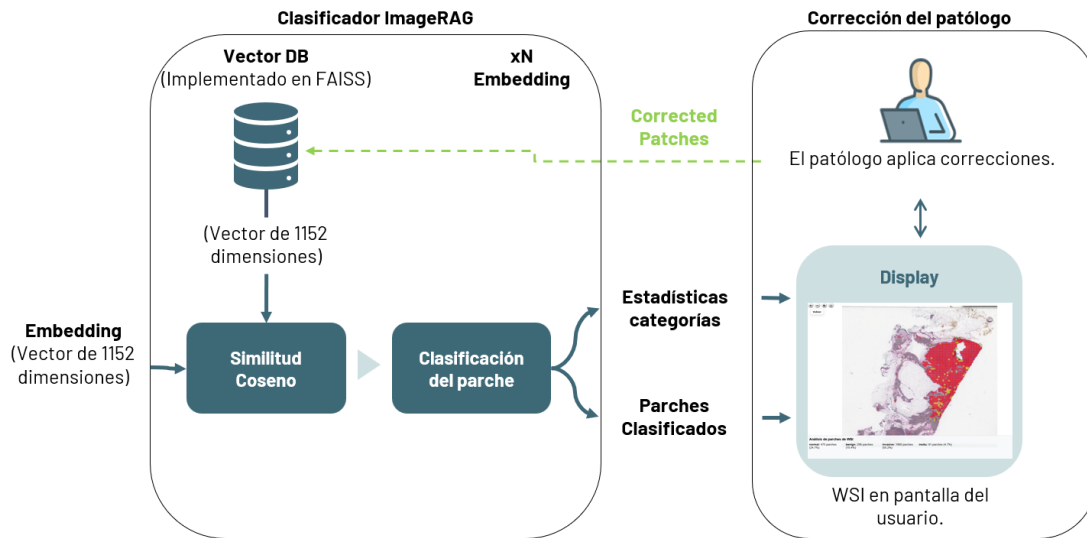


Figure 7.8: El usuario observa WSI, presiona un parche, lo corrige, y se almacena en feedback para eventual mejora del modelo.

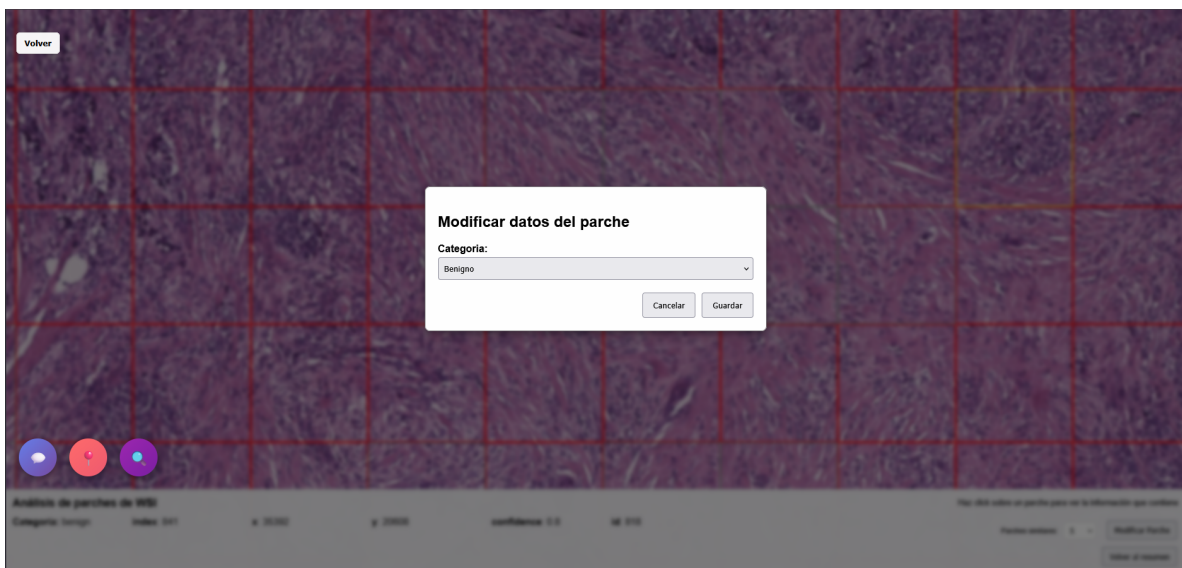


Figure 7.9: Interfaz de corrección de clasificación. El usuario al seleccionar un parche, tiene la posibilidad de modificar la clasificación de este, lo que provoca que el JSON con la información asociada a la WSI se actualice. Eventualmente, podemos extraer este parche y su nueva clasificación e incorporarlas al RAG visual.

Se implementó el flujo de retroalimentación explícita de la siguiente forma:

- **Acción del Experto:** El patólogo puede seleccionar un parche mal clasificado y asignar la etiqueta correcta mediante la UI.
- **Almacenamiento de Evidencia:** Al confirmar la corrección, el sistema extrae el parche de la WSI y lo almacena en un directorio de “feedback” junto con su nueva etiqueta.
- **Actualización de la Base de Datos Vectorial FAISS:** Eventualmente, estos parches corregidos son procesados por el encoder MedSigLIP e insertados en la base de datos FAISS, refinando así el espacio vectorial y mejorando la precisión futura del clasificador RAG.

Beneficios del diseño Human-in-the-Loop:

- **Mejora Continua:** El modelo se adapta a casos atípicos específicos
- **Validación por Experto:** Las correcciones crean un dataset anotado por patólogos reales.

La figura 7.10 muestra la interfaz de usuario implementada para este sistema. A la derecha se observa el cuadro que permite la comunicación con el sistema de agentes, mientras que en la parte inferior se observan estadísticas de la WSI clasificada.

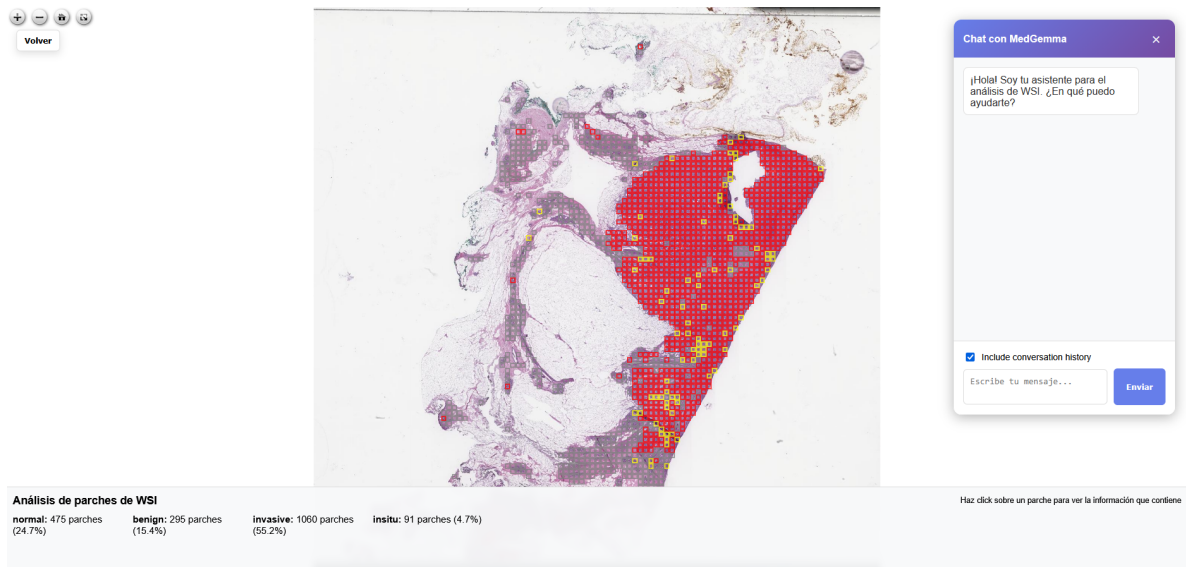


Figure 7.10: Interfaz de usuario del sistema. A la derecha se observa la ventana para interactuar con los agentes, en la parte inferior se observan las estadísticas de clasificación y en la parte central se observa la WSI clasificada.

3.7.2 Visualización de Parches Similares ImageRag

Para ver las imágenes que el clasificador ImageRAG considera similares en un parche para su clasificación, el usuario puede pedir una muestra de los vecinos más cercanos dentro de la base de datos vectorial. La figura 7.11 enseña la similitud del parche respecto a sus similares dentro de la base vectorial.

Parches Similares - ImageRAG

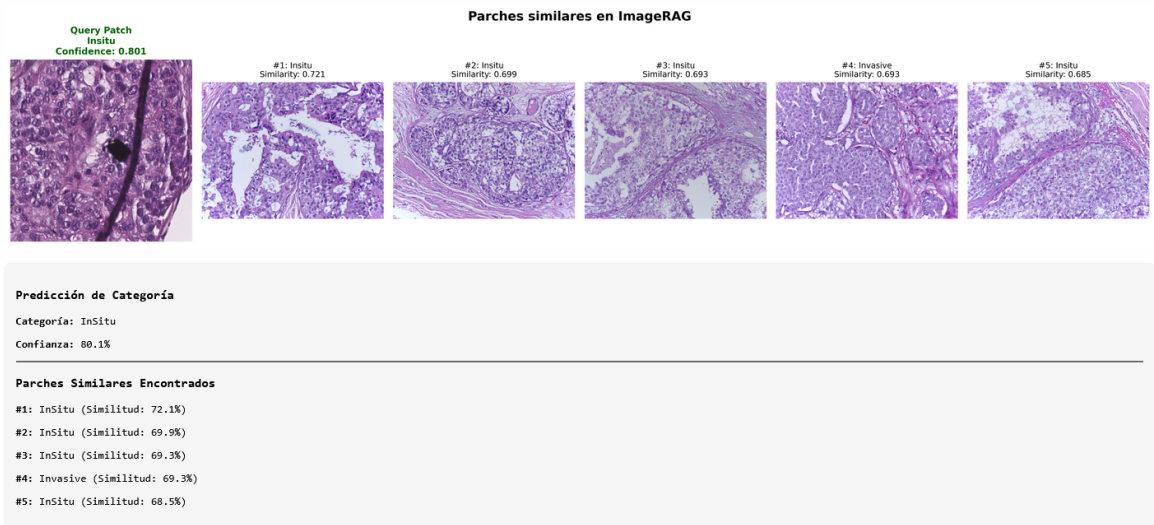


Figure 7.11: Visualización de imágenes similares al parche seleccionado. Esto permite al patólogo analizar la razón de la elección de categoría.

Para lograr este resultado, el sistema a través de la API, solicita obtener el parche seleccionado en la interfaz de usuario, recupera las imágenes similares de la base de datos vectorial, y construye una imagen comparativa por medio de **matplotlib**.

3.7.3 Comunicación entre Módulos

La arquitectura del sistema adopta un modelo **desacoplado entre Cliente-Servidor**, asegurando que existe una separación estricta entre la capa de presentación (Interfaz de Usuario) y la data procesada (backend). Esta separación permite un escalado independiente y facilita el mantenimiento.

- **Frontend (Cliente):** Aplicación web SPA (Single Page Application) que se ejecuta en el navegador del patólogo, responsable de visualización y captura de interacciones.
- **Backend (Servidor):** Servidor Python con FastAPI que orquesta los módulos de IA (Image RAG, LLM) y expone endpoints RESTful.

- **Protocolo Comunicación de Modelos:** Model Context Protocol (MCP) [14] encargado de estandarizar las comunicaciones entre la API y los agentes. Facilita la integración de modelos y herramientas en el servidor.

El protocolo de comunicación opera de la siguiente forma:

- **Request:** El frontend maneja peticiones HTTP asincrónicas (GET, POST, PUT, DELETE) usando la **browser's Fetch API**.
- **Payload:** La data es intercambiada en el formato **JSON** (JavaScript Object Notation). Por ejemplo, cuando el usuario consulta en el LLM, el frontend construye una estructura que contiene las conversaciones pasadas, información de la clasificación RAG y la pregunta, dentro de un objeto JSON.

Integración de Herramientas a via MCP

Una característica distinta en la integración del backend es el uso del **Model Context Protocol (MCP)** para aislar la lógica de los modelos de DeepLearning.

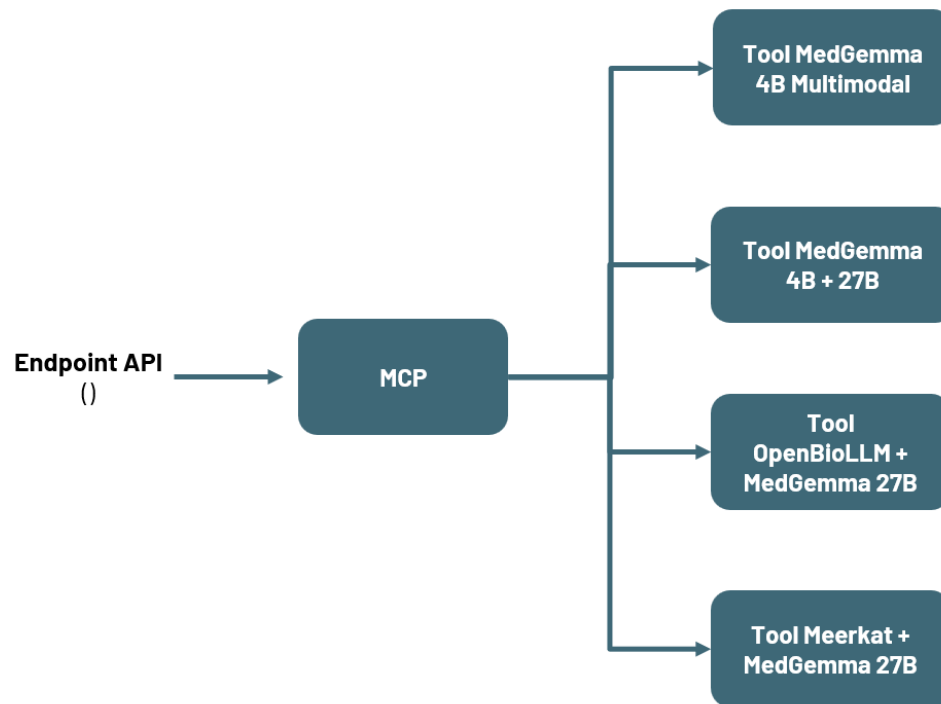


Figure 7.12: Diagrama sobre el funcionamiento del protocolo MCP. El módulo «MCP» orquesta las llamadas a las distintas herramientas. Las herramientas corresponden a rutinas de inferencia de los small model languages.

En vez de incorporar la lógica de los modelos directo al hilo principal de la API, el sistema usa el `stdio_client` para generar una herramienta «modelo LLM» que es un subproceso independiente.

- **Aislación de Procesos:** La API actúa como un orquestador. Cuando una petición de tipo `/chat` es recibida, la API establece una sesión con el script local de Python `./tools/*.py` via standard input/output (stdio) streams.
- **Estabilidad:** Esta arquitectura nos asegura que si la inferencia con el modelo LLM provoca un overflow de memoria o un timeout, nuestra página web encargada de manejar la visualización de la WSI no se terminará abruptamente.

3.7.4 Despliegue

La lógica de la aplicación está expuesta vía **ASGI (Asynchronous Server Gateway Interface)**, específicamente **Uvicorn**.

- **Concurrencia:** A diferencia de los tradicionales sincrónicos servers WSGI, Uvicorn hace uso de las capacidades de asincio. Esto permite al servidor manejar operaciones I/O no bloqueantes-como lo puede ser leer una wsi desde el disco de archivos del servidor o esperar la inferencia de un modelo LLM-sin tener que bloquear el hilo principal. Esto es crítico para mantener respuestas por parte del frontend mientras se están haciendo operaciones pesadas computacionalmente en el backend.

3.7.5 Limitaciones de la Interfaz de Usuario

Si bien la interfaz de usuario se diseñó pensando en su máxima optimización, existen decisiones de software que pueden perjudicar al usuario en ciertas ocasiones:

- **Utilización de Overlays:** Al observar los parches y su clasificación a través de overlays en HTML, permite que la programación sea intuitiva y de fácil mantención. Sin embargo, si el tamaño de los parches es menor (e.g., 256x256) el navegador sufre un peak de carga, que en equipos con poca disponibilidad de RAM provoca un fluido torpe por la WSI.
- **Carga de WSI:** El alto peso de las WSI (aprox. 2GB-5GB) es un problema a la hora de subirlas al servidor. Esto provoca que el software presente latencias en el flujo de ejecución (sin contar que una vez subidas las WSI, ocurre el proceso de clasificación)

Resultados y Análisis

Este capítulo presenta la evaluación sistemática de la prueba de concepto desarrollada. El análisis se divide en dos dimensiones críticas: la validación algorítmica, centrada en la precisión diagnóstica de los modelos de inteligencia artificial (clasificación de imágenes y orquestación de LLMs), y la validación sistémica, que examina la funcionalidad, latencia y capacidad de respuesta del software en un flujo de trabajo patológico simulado.

4.1 Definición de Pruebas y Métricas

Para cuantificar el desempeño de la herramienta, se seleccionaron métricas estándar en la literatura de ingeniería biomédica y sistemas distribuidos. A continuación, se detallan los criterios de evaluación:

4.1.1 Rendimiento del Clasificador (Image RAG)

Se midió el desempeño del clasificador Image RAG utilizando una estrategia de *k-fold cross-validation* ($k = 5$). Este método particiona el dataset de ICIAR 2018 en cinco subconjuntos distintos, entrenando y validando el modelo iterativamente. Se ha decidido utilizar 5 subconjuntos ya que una mayor cantidad provocará que los modelos sean más similares (y por consecuencia, menos independientes). Mientras menos independiente sean los modelos, estos tendrán mayor correlación, y el overfitting se esconderá.

Métricas seleccionadas:

- Exactitud (Accuracy) por clase
- Desviación estándar entre folds
- Confianza promedio del modelo
- Matriz de confusión agregada

4.1.2 Eficacia de la Orquestación LLM

Se evaluó cualitativamente el impacto de la arquitectura "Generador-Evaluador". Se compararán las respuestas generadas por 3 'small language models' frente a las refinadas por el modelo evaluador (MedGemma 27B cuantizado), determinando si el costo computacional adicional se traduce en una mejora clínica tangible.

Small Language Models:

- **MedGemma 4b-it:** Este modelo perteneciente a Google, presenta un entrenamiento base en el área médica, proporcionando mayor contexto a las respuestas y un mayor conocimiento del área a evaluar (patología).
- **Meerkat 7b:** Corresponde a un modelo instruction-tuned en medicina, que logró superar el 60% de aprobación en el «United States Medical Licensing Examination (USMLE)». El modelo fue entrenado usando un dataset sintético compuesto de razonamiento «Chain-Of-Thought» de alta calidad, extraído desde 18 textos médicos.
- **OpenBioLLM 8b:** Modelo de lenguaje diseñado específicamente para dominio médico. Desarrollado por Saama AI Labs, este modelo nivela técnicas modernas para alcanzar el estado del arte en un gran rango de tareas biomédicas.

Para evaluar el desempeño de la arquitectura, se seleccionaron 5 preguntas representativas de situaciones de duda por parte del patólogo ante situaciones inusuales, de forma

que el modelo reflexione antes de entregar una respuesta.

Preguntas seleccionadas:

Pregunta 1 I am staging a breast excision case with extensive high-grade DCIS. I found three separate foci of microinvasion. Focus A measures 0.8 mm, Focus B measures 1.2 mm, and Focus C measures 0.5 mm. Should I add these diameters together to determine the T-stage, or do I use the largest one? What is the correct pT category for this patient?

Fuente: AJCC Cancer Staging Manual, 8th Edition (Breast Chapter).

Pregunta 2 I need to calculate the TILs (Tumor Infiltrating Lymphocytes) score for a Triple Negative Breast Cancer case to report the prognostic value. I see a lot of lymphocytes in the stroma, but also many infiltrating the tumor nests themselves, and a dense cluster inside a central area of comedo-necrosis. According to the International Immuno-Oncology Biomarker Working Group guidelines, which of these specific compartments do I include in the final percentage score?

Fuente: Salgado R, et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014.

Pregunta 3 I am looking at a lesion composed of small, open, round glands infiltrating the adipose tissue without a stromal desmoplastic reaction. The glands contain eosinophilic colloid-like secretion. It looks like Tubular Carcinoma, but the glands are S100 positive and ER negative. Crucially, if I stain for Laminin or Collagen IV, what finding would confirm this is Microglandular Adenosis (MGA) rather than Tubular Carcinoma, considering MGA lacks myoepithelial cells?

Fuente: Clement PB, Azzopardi JG. Microglandular adenosis of the breast—a lesion simulating tubular carcinoma.

Pregunta 4 I have a core biopsy showing a cribriform proliferation of monotonous cells. The

architecture is rigid, but the lesion only partially involves two duct spaces and measures exactly 1.5 mm in greatest contiguous dimension. According to the WHO Classification (5th Edition), is this sufficient for a diagnosis of Ductal Carcinoma In Situ (DCIS), or should it be downgraded?

Fuente: WHO Classification of Tumours: Breast Tumours (5th Edition).

Pregunta 5 I'm analyzing a cribriform lesion that looks like low-grade DCIS. However, the 'lumens' are filled with a fibrillar, slightly basophilic material rather than secretion. If these 'spaces' are actually enclosed by myoepithelial cells rather than luminal epithelial cells, what is the specific diagnosis, and what material makes up the 'spherules'?

Fuente: Grin A, et al. Collagenous Spherulosis of the Breast.

Los criterios cualitativos usados para evaluar la respuesta de los modelos, se basan en el RACCCA framework. Diseñado por Andrew Maynard, este framework entrega 6 criterios de evaluación para determinar la calidad de las respuestas:

- Relevance (pertinencia): Hasta que punto la respuesta aborda el problema o la pregunta directamente.
- Accuracy (precisión): El grado en que la respuesta provee información correcta, confiable y basada en hechos.
- Appropriateness (idoneidad): Que tan bien la respuesta se adecúa a la audiencia objetivo, mientras que es respetuoso en el tono y contenido.
- Completeness (integridad): El grado en que la respuesta cubre los aspectos esenciales de la pregunta.
- Coherence (coherencia): Que la respuesta se encuentre lógicamente estructurada y bien organizada.
- Clarity (claridad): Que tan fácil la respuesta puede ser entendida por la audiencia

objetivo.

4.1.3 Latencia del Sistema

Se midió el tiempo de respuesta "extremo a extremo" (end-to-end) para operaciones críticas:

- Clasificación de una WSI usando ImageRAG para diferentes tamaños de parches
- Recuperación de parches similares mediante RAG visual
- Generación de respuesta de agentes LLM

Respecto a lo que se considera una latencia deseable, no se hallaron documentos legales o de instituciones relacionadas al software que establezcan un tiempo límite. Por lo mismo, queda a criterio de la institución o del cliente que eventualmente adquiriría esta prueba de concepto esa elección. Sin embargo, ya que se planea usar en situaciones «cotidianas», lo ideal es que el sistema pueda funcionar en equipos con capacidad de cómputo limitada (para el frontend, ya que el servidor se hará cargo del cómputo pesado como la clasificación de la WSI).

4.2 Resultados Experimentales

4.2.1 Desempeño del Clasificador Image RAG

La tabla 2.1 resume las métricas de rendimiento promedio obtenidas tras la validación cruzada de 5 folds.

Podemos observar como es que la exactitud para las categorías Normal e Invasive fue mayor al 90%, siendo estas de especial relevancia para un diagnóstico.

En la tabla 2.2 podemos observar comparar los resultados que obtuvieron los diferentes equipos en el desafío ICIAR, respecto a nuestro trabajo [3]:

Table 2.1: Métricas de desempeño del clasificador Image RAG por clase (Promedio de 5-fold CV).

Categoría	Exactitud	Desv. Est.	Conf. Prom.	Conf. Std
Normal	0.99	± 0.02	0.869	0.010
Benign	0.72	± 0.05	0.648	0.030
InSitu	0.84	± 0.02	0.783	0.046
Invasive	0.95	± 0.03	0.914	0.009

Nota: La exactitud global media del sistema fue de 0.875 ± 0.017 .

Table 2.2: Comparación de ImageRAG respecto a trabajos presentados al concurso ICIAR. Acc. corresponde al promedio de precisión obtenido para cada clase (para ImageRAG, corresponde al promedio de los 5 folds), mientras que Input Size corresponde al tamaño de los parches utilizados para clasificar.

Team	Acc.	Approach	Input Size
ImageRAG (nuestro)	0.875	MedSigLIP + RAG	448×448
(Chennamsetty et al., 2018)	0.87	Resnet-101; Densenet-161	224×224
(Kwok, 2018)	0.87	Inception-Resnet-v2	299×299
(Brancati et al., 2018)	0.86	Resnet-34, 50, 101	308×308
(Marami et al., 2018)	0.84	Inception-v3	512×512
(Kohl et al., 2018)	0.83	Densenet-161	205×154
(Wang et al., 2018a)	0.83	VGG16	224×224
Steinfeldt et al.	0.81	XCception	229×229
(Kon ´e and Boulmane, 2018)	0.81	ResNeXt50	299×299
Nedjar et al.	0.81	Inception-v3, Resnet-50, Mobile-Net	224×224

Nuestro trabajo presenta un desempeño similar al ganador de aquel entonces, pero debemos considerar que nuestra solución implementa métodos más modernos a los

usados por aquel entonces (2018). Sin embargo, no deja de ser relevante una precisión cercana al 88%, con la posibilidad de observar los parches similares (interpretabilidad).

4.2.1.1 Análisis de la Matriz de Confusión

La Figura 2.1 presenta la matriz de confusión agregada del clasificador Image RAG, obtenida al consolidar las predicciones de los 5 folds de validación cruzada.

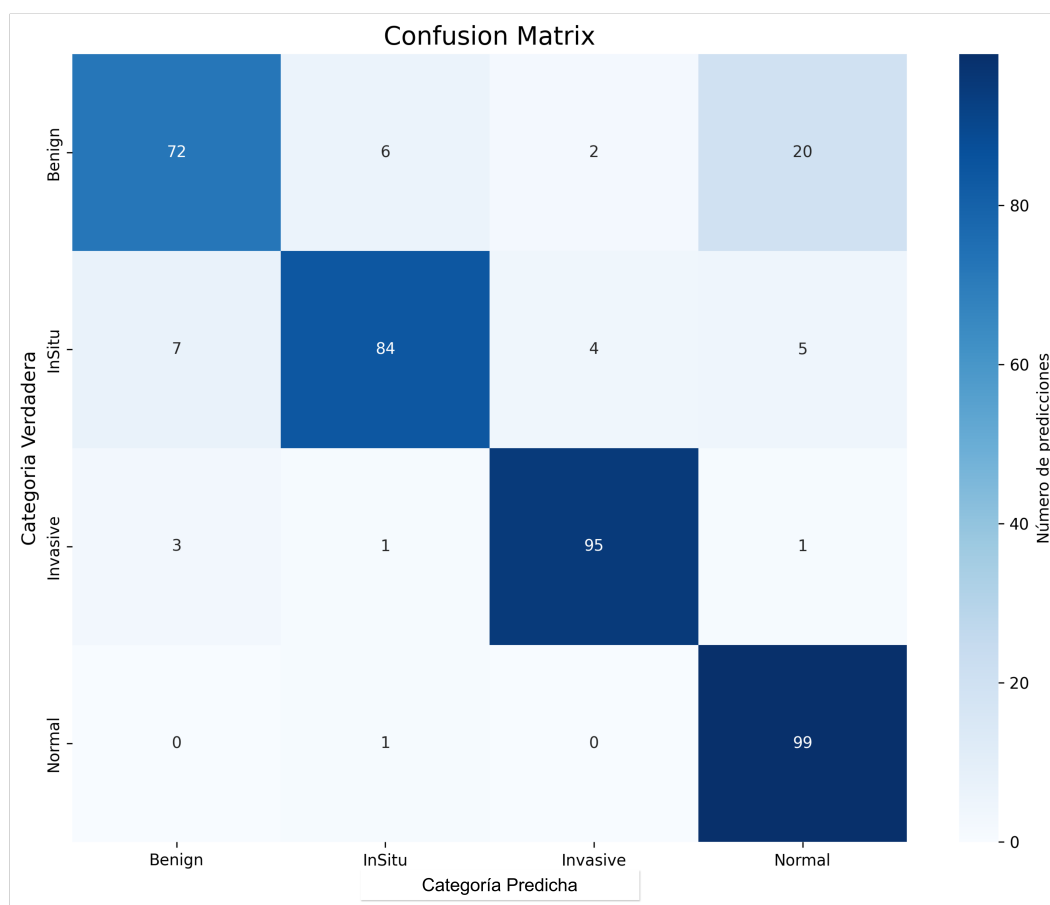


Figure 2.1: Matriz de confusión del Image RAG.

La tabla 2.3 contiene las métricas de la matriz de confusión:

A partir de la matriz de confusión, se calculan las siguientes métricas de desempeño:

Precisión por clase:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (2.1)$$

Table 2.3: Métricas de la matriz de confusión por clase del clasificador ImageRAG.

Clase	TP	FP	FN	TN
Benign	72	10	28	290
InSitu	84	8	16	292
Invasive	95	6	5	294
Normal	99	26	1	274

Nota: TP = True Positive, FP = False Positive, FN = False Negative, TN = True Negative.

Recall/Sensibilidad por clase:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \quad (2.2)$$

F1-Score por clase:

$$F1_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (2.3)$$

Métricas Macro-Promedio:

$$\text{Macro-Precision} = \frac{1}{N} \sum_{i=1}^N \text{Precision}_i = 0.88 \quad (2.4)$$

$$\text{Macro-Recall} = \frac{1}{N} \sum_{i=1}^N \text{Recall}_i = 0.88 \quad (2.5)$$

$$\text{Macro-F1} = \frac{1}{N} \sum_{i=1}^N F1_i = 0.87 \quad (2.6)$$

donde $N = 4$ es el número de clases. La tabla 2.4 resume los resultados obtenidos:

Table 2.4: Métricas de desempeño del clasificador ImageRAG por clase.

Clase	Precisión	Recall	F1-Score
Benign	0.87	0.72	0.79
InSitu	0.91	0.84	0.87
Invasive	0.94	0.95	0.94
Normal	0.79	0.99	0.88
Macro-Promedio	0.88	0.88	0.87

El clasificador desarrollado demuestra un desempeño robusto en la tarea de clasificación multiclase, alcanzando métricas macro-promediadas de 0.88 en precisión y recall, y 0.87 en F1-score. Estos resultados indican que el modelo logra un balance adecuado entre la capacidad de identificar correctamente cada clase y minimizar las clasificaciones erróneas.

La clase Invasive presenta el mejor desempeño global con un F1-score de 0.94, combinando alta precisión (0.94) y recall (0.95), lo que sugiere que el modelo captura eficientemente los patrones distintivos del cáncer invasivo con mínimos falsos positivos y falsos negativos. La clase Normal exhibe un recall excepcional de 0.99, indicando que el clasificador raramente clasifica incorrectamente tejido normal, una característica crítica para evitar sobre-diagnósticos. Sin embargo, su precisión más baja (0.79) revela que el modelo tiende a clasificar algunas muestras de otras clases como Normal, lo cual está relacionado con la similitud morfológica entre tejido normal y benignas.

Las clases InSitu y Benign muestran un desempeño intermedio. Para InSitu, el F1-score de 0.87 refleja un balance razonable, aunque los 16 falsos negativos sugieren que algunas lesiones in situ podrían estar siendo confundidas con otras categorías. La clase Benign presenta el mayor desafío para el clasificador, con el F1-score más bajo (0.79) y 28 falsos negativos, lo que indica dificultad para distinguir lesiones benignas de otras clases, particularmente de tejido normal o lesiones in situ, probablemente debido a la variabilidad morfológica inherente a esta categoría. [33] El estudio de «Quinn C, Maguire A, Rakha E. Pitfalls in breast pathology» reconoce que la similitud entre casos benignos y casos normales e invasivos, es un problema recurrente en la interpretación de tejidos teñidos con H&E.

Desde una perspectiva clínica, estos resultados son prometedores. La alta sensibilidad en la detección de casos Normal (0.99) y el sólido desempeño en Invasive (0.95) son especialmente relevantes, ya que minimizan el riesgo de pasar por alto patologías malignas. No obstante, el menor recall en Benign (0.72) representa un área de mejora importante.

La figura 2.2 representa un resumen de los resultados obtenidos para cada subconjunto de datos. Se aprecia como es que las categorías de Benign e InSitu presentan un desempeño inferior respecto a las categorías Normal e Invasive. Finalmente, la precisión media del clasificador Image Rag para los subconjuntos es de 87.5%.

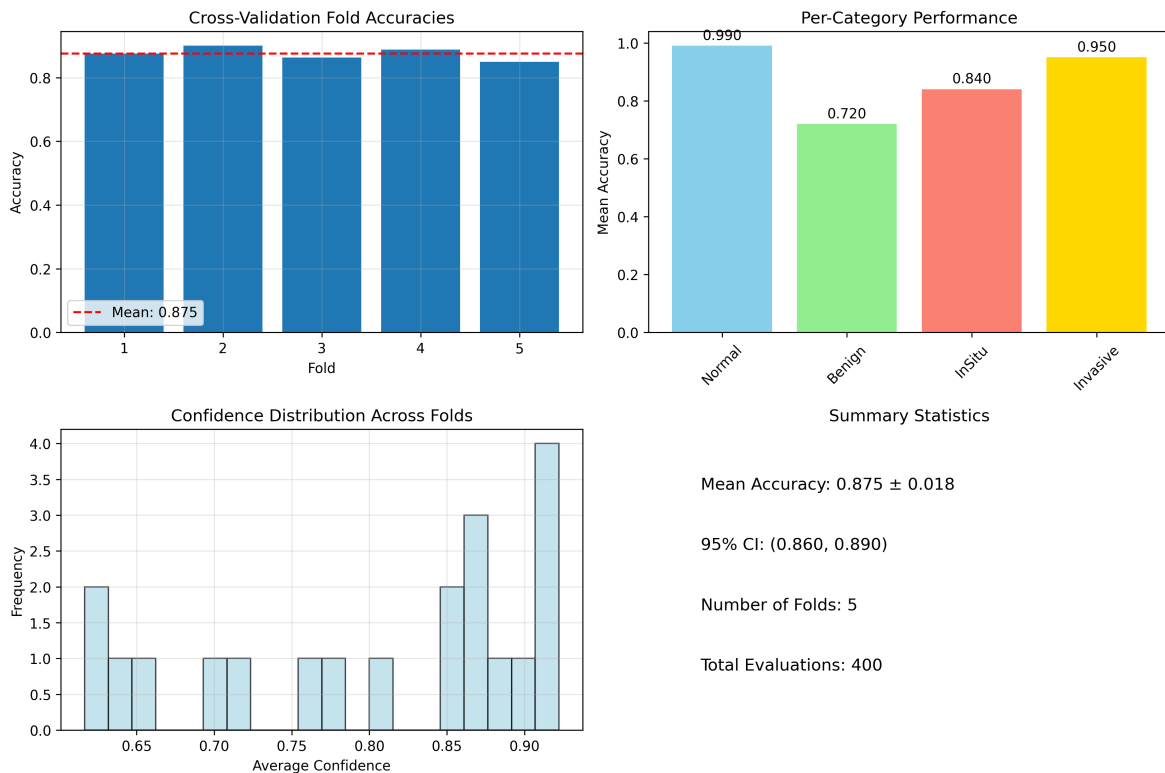


Figure 2.2: Resultados k-fold cross validation para clasificador Image RAG.

4.2.2 Evaluación de la Orquestación LLM

Debido a que no se contó con el apoyo de patólogos para la evaluación de las respuesta, se optó por utilizar un LLM comercial como evaluador. El modelo elegido corresponde a «Claude Sonnet 4.5», y el prompt utilizado se encuentra en el apéndice C 5.5

Para facilitar la lectura de los resultados obtenidos, continuación se presentará la respuesta esperada por los modelos, la respuesta del modelo MedGemma 4B y finalmente un cuadro comparativo de los puntajes obtenidos para cada modelo.

Respuesta Esperada Pregunta 1

Do NOT sum. The stage is based on the largest single focus.
 Since Focus B is 1.2 mm, it exceeds the limit for microinvasion (T1mi is ≤ 1.0 mm). The correct stage is pT1a (tumors >1 mm and ≤ 5 mm).

The presence of **multiple** foci should be noted with the (m) modifier (e.g., pT1a(m)).

MedGemma 4B

Respuesta Inicial:

2 The AJCC TNM staging system for breast cancer uses the ***largest*** dimension of the tumor to determine the T-stage. In this case, the largest focus of microinvasion is Focus B, measuring 1.2 mm. Therefore, the correct pT category for this patient is pT2.

Respuesta Final:

The AJCC TNM staging system for breast cancer determines the T-stage based on the size of the largest invasive tumor focus. You should not add the diameters together.

7 In this case, you have three separate foci of microinvasion measuring 0.8 mm, 1.2 mm, and 0.5 mm. The largest of these is 1.2 mm.

Based on the AJCC staging criteria:

pT1: Tumor ≤ 20 mm

pT1a: Tumor ≤ 5 mm

pT1b: Tumor > 5 mm but ≤ 10 mm

12 pT1c: Tumor > 10 mm but ≤ 20 mm

Since the largest focus measures 1.2 mm, the correct pT category **for** this patient is pT1b.

	Relevancia	Precisión	Idoneidad	Integridad	Coherencia	Claridad
MedGemma 4B	4	2	5	3	5	5
Meerkat 7B	4	2	4	2	4	4
OpenBioLLM 8B	5	2	4	2	5	4

Respuesta Esperada Pregunta 2

You count Stromal TILs only. You Exclude Intratumoral TILs (lymphocytes in direct contact with tumor cells without intervening stroma). You Exclude TILs in areas of necrosis or hyalinization. The score is defined as the \% of stromal area occupied by mononuclear cells.

MedGemma 4B Respuesta

Respuesta Inicial: Based on the provided information, the key points for calculating TILs in Triple Negative Breast Cancer are:

Compartments to include: Lymphocytes in the stroma and infiltrating the tumor nests.

4 Exclude: Carcinoma cells and peritumoural follicular aggregates and tertiary lymphoid structures with germinal centres.

Methodology: Quantify TILs on H&E-stained tissue sections at 20-40x magnification, on the most representative tumor block. Score in the stroma between the areas of carcinoma.

Reporting: Report as a percentage, based on the available tissue analysed. If heterogeneous, report an average, disregarding hotspots.

Therefore, the specific compartments to include in the final percentage score are the lymphocytes in the stroma and infiltrating the tumor nests.

9 Respuesta Final: Calculating the Tumor Infiltrating Lymphocytes (TILs) score is crucial for assessing prognosis in Triple Negative Breast Cancer (TNBC). Here's how to determine the score based on standard recommendations:

Compartments to Include:

14 Stromal TILs: Lymphocytes located in the stroma, which is the

connective tissue between the tumor nests.

Intratumoral TILs: Lymphocytes that have infiltrated inside the tumor nests themselves.

Compartments to Exclude:

19 Carcinoma Cells: The score is a percentage of the stromal area, so carcinoma cells are not included in the denominator.

Peritumoural Follicular Aggregates and Tertiary Lymphoid Structures: Lymphocytes forming distinct structures like follicles or tertiary lymphoid structures (often found near the tumor but separate from the stroma or tumor nests) are not included in the stromal TIL assessment.

Lymphocytes within Comedo-Necrosis: Lymphocytes within the central area of comedo-necrosis should be assessed based on their location - if they are infiltrating the tumor nests, they are included as intratumoral TILs. If they are within the stroma surrounding the necrosis, they are included as stromal TILs.

Methodology for Scoring:

24 Tissue: Use H&E-stained tissue sections.

Magnification: Examine at 20-40x magnification with a 10x ocular.

Sample: Use the most representative tumor block available (biopsy or surgical specimen).

Assessment: Score the percentage of lymphocytes (including both lymphocytes and plasma cells) within the stromal areas between the tumor nests.

29 Reporting: Report the final score as a percentage. If the distribution of TILs is heterogeneous, calculate an average score across the assessed tissue, disregarding any "hotspots" (areas with exceptionally high density).

Therefore, for your case, you should include both the lymphocytes infiltrating the tumor nests and those present in the stroma

between the nests when calculating the final percentage score.

	Relevancia	Precisión	Idoneidad	Integridad	Coherencia	Claridad
MedGemma 4B	3	1	4	2	4	4
Meerkat 7B	4	2	5	3	5	5
OpenBioLLM 8B	4	1	5	2	5	4

Respuesta Esperada Pregunta 3

MGA will show a thick, intact basement membrane around the glands (positive for Laminin/Collagen IV).

Usually, invasive carcinomas lack a basement membrane, and benign glands have myoepithelial cells. MGA is the rare exception that lacks myoepithelial cells (like cancer) but retains the basement membrane (unlike cancer).

MedGemma 4B Respuesta

Respuesta Inicial: The key to differentiating Microglandular Adenosis (MGA) from Tubular Carcinoma in this scenario is the absence of myoepithelial cells. Here's why Laminin or Collagen IV staining would be helpful:

MGA: MGA is characterized by the absence of myoepithelial cells. Laminin and Collagen IV are components of the basement membrane, which surrounds the glands. If these stains are negative for myoepithelial cells, it strongly supports the diagnosis of MGA.

Tubular Carcinoma: Tubular Carcinoma, by definition, does have myoepithelial cells. Therefore, staining for Laminin and Collagen IV would likely show the presence of myoepithelial cells, which would be a key finding supporting the diagnosis of Tubular Carcinoma.

7 Therefore, the finding that would confirm MGA rather than Tubular Carcinoma is the absence of myoepithelial cells when staining for Laminin or Collagen IV.

Respuesta Final: The key to differentiating Microglandular Adenosis (MGA) from Tubular Carcinoma using Laminin or Collagen IV staining lies in the presence or absence of myoepithelial cells surrounding the glands.

Microglandular Adenosis (MGA): MGA is characterized by the absence of myoepithelial cells. Laminin and Collagen IV are components of the basement membrane surrounding the glands. Therefore, staining for Laminin or Collagen IV in MGA will highlight the basement membrane outlining the small, round glands, but there will be no staining of myoepithelial cells around the glands.

12 Tubular Carcinoma: Tubular Carcinoma typically has myoepithelial cells surrounding the tubules. Therefore, staining for Laminin or Collagen IV would show the basement membrane outlining the tubules and positive staining of myoepithelial cells surrounding them.

Thus, the finding that would confirm MGA rather than Tubular Carcinoma is the absence of myoepithelial cells when staining for Laminin or Collagen IV around the glands.

	Relevancia	Precisión	Idoneidad	Integridad	Coherencia	Claridad
MedGemma 4B	1	0	1	1	3	4
Meerkat 7B	2	1	4	1	4	4
OpenBioLLM 8B	4	0	4	1	5	5

Respuesta Esperada Pregunta 4

It should be downgraded to Atypical Ductal Hyperplasia (ADH). The Rule : Low-grade cribriform proliferations are classified as DCIS only if they exceed 2 mm in size or completely involve >2 duct spaces. Lesions smaller than this (1.5 mm) are diagnostic of ADH.

MedGemma 4B Respuesta

Respuesta Inicial: Based on the provided information, the diagnosis of Ductal Carcinoma In Situ (DCIS) is a possibility, but further evaluation is needed. Here's a breakdown:

Key Findings: Cribriform proliferation of monotonous cells, rigid architecture, partial involvement of two duct spaces, and a dimension of 1.5 mm.

Relevant Information from the RAG: The RAG information from the WHO text discusses atypical ductal hyperplasia (ADH) and its relationship to DCIS. It highlights that lesions of limited extent, like the one described, are often classified as ADH on core needle biopsy. The WHO recommends a conservative approach in such cases, awaiting clinical trial results. The RAG also mentions that ADH represents a spectrum of atypical architectural patterns, ranging from flat epithelial atypia to low-grade DCIS.

Analysis: The description of the lesion (cribriform proliferation, monotonous cells, rigid architecture, partial ductal involvement, and small dimension) is consistent with ADH. The fact that the lesion only partially involves two duct spaces and measures 1.5 mm in greatest contiguous dimension is also consistent with ADH.

Conclusion: While the findings are suggestive of DCIS, the limited extent of the lesion (1.5 mm) makes it more likely to be classified as ADH on core needle biopsy. The WHO recommends a conservative approach in such cases, awaiting clinical trial results.

Therefore, the answer is: It is possible that it is sufficient for a diagnosis of Ductal Carcinoma In Situ (DCIS), but it should be downgraded to atypical ductal hyperplasia (ADH) on core needle biopsy due to the limited extent of the lesion.

Respuesta Final: Based on the description provided, the lesion is best classified as ****Atypical Ductal Hyperplasia (ADH)**** on the core needle biopsy.

14

Here's the reasoning:

DCIS vs. ADH: Ductal Carcinoma In Situ (DCIS) involves a more extensive proliferation of malignant cells within the duct system. Atypical Ductal Hyperplasia (ADH) is characterized by atypical cells but involves a more limited proliferation.

Size Criteria: The lesion described is very small (1.5 mm) and only partially involves two duct spaces. Diagnostic criteria often use a threshold of 2 mm for distinguishing ADH from low-grade DCIS, especially in core biopsies.

19

Core Biopsy Limitations: As highlighted in the provided reference material, core needle biopsies may only sample a small portion of the overall lesion. Therefore, a conservative diagnosis is often preferred when the extent is limited. A lesion measuring 1.5 mm is typically classified as ADH in this setting.

Cellular Features: While cribriform proliferation and monotonous cells can be seen in low-grade DCIS, their presence within such a limited extent (1.5 mm) strongly favors ADH on a core biopsy.

Therefore, while the cellular features might raise suspicion for DCIS, the limited extent of the lesion (1.5 mm) makes ADH the appropriate diagnosis for this core biopsy sample. This diagnosis may influence further management, often involving correlation with imaging and potentially surgical excision for definitive diagnosis and staging.

	Relevancia	Precisión	Idoneidad	Integridad	Coherencia	Claridad
MedGemma 4B	5	5	5	5	5	5
Meerkat 7B	5	5	5	2	3	3
OpenBioLLM 8B	5	5	5	1	3	4

Respuesta Esperada Pregunta 5

Diagnosis: Collagenous Spherulosis.

Material: The spherules are composed of Basement Membrane material (Collagen IV), not mucin.

3 Distinction: In DCIS, the cribriform spaces are true lumens lined by epithelial cells. In Collagenous Spherulosis, the material is extracellular matrix produced by myoepithelial cells.

MedGemma 4B Respuesta

2 Respuesta Inicial: Based on the provided information, the most likely diagnosis is cribriform ductal carcinoma in situ (DCIS).

The key features that point to this diagnosis are:

Cribriform pattern: The lesion is described as "cribriform," which is a characteristic architectural pattern of DCIS.

Low-grade DCIS: The lesion is described as "low-grade," indicating a less aggressive form of DCIS.

7 Fibrillar, slightly basophilic material in the lumens: This is a crucial finding. The description of the material filling the spaces as "fibrillar, slightly basophilic" is suggestive of **myoepithelial cells** rather than the typical secretion seen in normal or low-grade DCIS.

If the spaces are indeed enclosed by myoepithelial cells, this would be a specific type of cribriform DCIS.



Regarding the "spherules," the material filling the spaces is likely myoepithelial cells. Myoepithelial cells are specialized cells that surround the ducts in the breast and are responsible for contraction. In DCIS, these cells can become abnormal and form spherules or other structures within the lumen.

12

Therefore, the specific diagnosis is cribriform DCIS with myoepithelial cell spherules.

Respuesta Final: Based on your description, the most likely diagnosis is **Cribriform Ductal Carcinoma In Situ (DCIS)**.

17

Here's a breakdown:

1. **Cribriform Pattern:** This refers to the architectural arrangement where epithelial cells form gland-like structures or spaces within the ducts.
2. **Myoepithelial Cells:** The crucial finding is that the spaces are enclosed by myoepithelial cells rather than luminal epithelial cells. This confirms the lesion is *in situ* (confined to the ducts) and not invasive. The fibrillar, slightly basophilic material filling these spaces is consistent with myoepithelial cells.
3. **Material/Spherules:** The "spherules" you are observing are likely clusters or aggregates of these myoepithelial cells within the lumen of the duct.
- 22 4. **Grading:** DCIS is graded based on nuclear cytology (low, intermediate, or high grade). Your initial assessment suggests low-grade DCIS, which typically features small, uniform nuclei. Definitive grading requires careful microscopic evaluation of the cell nuclei.

Therefore, the specific diagnosis is **Cribriform DCIS**, characterized by a cribriform pattern and the presence of myoepithelial cells enclosing the spaces, with the material within the spaces being these myoepithelial cells (forming spherules or

clusters). The grade would be determined by the nuclear features.

	Relevancia	Precisión	Idoneidad	Integridad	Coherencia	Claridad
MedGemma 4B	1	0	2	1	3	4
Meerkat 7B	1	1	1	0	2	1
OpenBioLLM 8B	1	0	1	0	2	1

Los tres modelos evaluados (MedGemma, Meerkat y OpenBioLLM) presentaron patrones de desempeño similares a lo largo de las pruebas. La principal mejora observada en la mayoría de los casos correspondió a la claridad y estructura de las respuestas generadas por el modelo evaluador. Sin embargo, un aspecto particularmente preocupante fue la baja precisión de las respuestas, especialmente evidente en las preguntas 1, 2, 3 y 5, donde los modelos iniciales proporcionaron respuestas incorrectas que el modelo evaluador no logró detectar ni corregir.

Un caso destacable fue la pregunta 1 con el modelo Meerkat, donde el evaluador logró corregir un error crítico en el método de cálculo; no obstante, la respuesta final permaneció incorrecta a pesar de contar con todo el contexto necesario para resolver el problema adecuadamente. Este hallazgo sugiere una capacidad limitada del modelo evaluador para reflexionar sobre problemas matemáticos y lógicos complejos.

Por el contrario, la pregunta 4 mostró resultados positivos para todos los modelos. MedGemma alcanzó un desempeño perfecto, mientras que Meerkat y OpenBioLLM demostraron excelente precisión. Sin embargo, en estos dos últimos, el modelo evaluador eliminó reflexiones y justificaciones que podrían ser valiosas en un contexto médico, priorizando la brevedad sobre el contexto clínico. Un aspecto crucial de esta pregunta fue que el RAG textual proporcionó contexto clave para responder correctamente, lo que revela dos conclusiones importantes:

- El RAG textual es capaz de encontrar información relevante cuando está disponible en el corpus.

- El corpus de texto actual, aunque seleccionado con literatura relevante para el contexto de este trabajo, resultó insuficiente para las preguntas que requieren mayor reflexión por parte de los modelos. Para las preguntas 1, 2, 3 y 5, donde los modelos no alcanzaron la precisión necesaria, no se encontró contexto relevante en el RAG.

En conclusión, la arquitectura de agentes implementada no generó una mejora significativa en la calidad de las respuestas bajo las condiciones actuales. Sin embargo, no es posible determinar si un RAG textual más amplio produciría resultados diferentes. Los hallazgos sugieren que la limitación principal no reside en la arquitectura en sí, sino en la insuficiencia del corpus textual para proporcionar el contexto necesario al modelo generador.

Respecto a preguntas de «conocimiento general», el sistema de agentes tuvo un desempeño competente, ya que estas preguntas no requieren que los modelos realicen una reflexión profunda, además de encontrarse las respuestas en la información recuperada por el RAG textual (ya que es información general del contexto médico).

4.3 Rendimiento del Sistema y Validación Técnica

Esta sección evalúa el desempeño computacional del sistema completo, midiendo latencia en operaciones críticas del flujo de trabajo. Las métricas se obtuvieron en el hardware descrito en la subsección 3.1, ejecutando el sistema bajo condiciones de carga típicas de un entorno de diagnóstico.

4.3.1 Clasificación de una WSI usando ImageRAG para diferentes tamaños de parches

Para evaluar la escalabilidad del clasificador Image RAG, se procesó una WSI representativa (detallado en Anexo B 5.5) con tres configuraciones de tamaño de parche. La Tabla 3.5 resume los tiempos de procesamiento end-to-end:

Table 3.5: Latencia de clasificación de ImageRAG según tamaño de parche.

Tamaño de Parche (px)	Latencia (min:s)
256×256	8:06
448×448	3:56
800×800	2:25

Nota: Los tiempos incluyen segmentación Otsu, filtrado de calidad y clasificación K-NN. El throughput constante de ~ 5 parches/s refleja el cuello de botella en generación de embeddings (MedSigLIP en GPU).

A medida que el tamaño de los parches es menor, la latencia de clasificación aumenta debido a que la cantidad de parches aumenta. El clasificador ImageRAG tiene una tasa de clasificación de aproximadamente 5 parches por segundo, de manera que si la WSI presenta una mayor/menor presencia de parches válidos (es decir, que presente tejido) los tiempos presentarán variaciones significativas.

4.3.2 Recuperación de parches similares (múltiples vecinos)

Durante la interacción del patólogo con la interfaz, la búsqueda de casos similares debe ser suficientemente rápida para no interrumpir el flujo diagnóstico. Se midió la latencia de recuperación considerando el pipeline completo: extracción del parche desde la WSI, generación de embedding y búsqueda K-NN en FAISS. Los resultados se encuentran en la tabla 3.6

Table 3.6: Latencia de Recuperación de Parches Similares.

Número de Vecinos	Latencia (s)
5	7
10	8
15	12

Nota: El incremento desproporcionado en $k = 15$ se debe a la transferencia y renderizado de imágenes adicionales en el frontend, no al cómputo FAISS en sí.

Mientras la cantidad de vecinos aumenta, mayor es la latencia. Aunque definitivamente no presenta un alza de tiempo considerable.

Se han seleccionado vecinos de 5, 10 y 15 ya que son las opciones permitidas por nuestro modelo. No se seleccionó un número mayor de vecinos ya que no aportan información relevante (ya que dejan de ser similares a nuestro parche).

4.3.3 Generación de respuesta de agentes LLM (Múltiples modelos)

La latencia de generación de respuestas conversacionales es crítica para la percepción de interactividad del sistema. Se evaluaron las tres configuraciones de modelos generadores contra el evaluador fijo (MedGemma-27B).

Para medir la latencia en la generación de la respuesta de los agentes LLM, se utilizó la pregunta «What is a WSI?». De esta manera obtenemos una evaluación sistemática. En la tabla 3.7 se encuentran los resultados para cada agente:

Table 3.7: Latencia de respuesta del sistema multi-agente LLM.

Modelo	Parámetros	Gen. (s)	Eval. (s)	Total (s)
MedGemma	4B	47	192	239
Meerkat	7B	9	67	76
OpenBioLLM	8B	7	72	79

Nota: Gen. = Small Model, Eval. = Evaluación con MedGemma 27B. Meerkat 7B muestra la mejor relación eficiencia/tamaño ($3.1\times$ más rápido que MedGemma 4B).

La configuración de agentes que más tarda en ejecutarse corresponde a MedGemma 4B, mientras que la configuración que tarda menos corresponde a Meerkat 7B. La razón de estos resultados son asociados a las diferentes arquitecturas que presentan los modelos generadores, mientras que la complejidad y longitud de la respuesta del modelo generador también puede contribuir (ver respuestas del modelo MedGemma 4B en ??).

Si bien no se encontraron referencias a tiempos de ejecución óptimos o exigidos para aplicaciones médicas, las latencias presentes en la etapa de clasificación con ImageRAG (3.5), junto a las latencias obtenidas en la generación de respuestas (3.7) pueden ser perjudiciales si es que los exámenes deben ser analizados rápidamente. De todas maneras,

estos resultados se encuentran fuertemente limitados a nuestra capacidad de cómputo. Un equipamiento moderno, o un levantamiento de este sistema en un servidor dedicado, pueden favorecer la obtención de latencias aceptables para la aplicación propuesta.

El sistema demuestra latencias factibles para procesamiento batch (clasificación de WSI) pero requiere optimización sustancial para interacción conversacional en tiempo real. El cuello de botella principal reside en la fase de evaluación LLM (67-192s), que aunque costosa computacionalmente, es esencial para mitigar hallucinations y garantizar fiabilidad médica.

Conclusiones y Trabajo Futuro

Este capítulo sintetiza los hallazgos principales de la investigación, evaluando el cumplimiento de los objetivos planteados y analizando críticamente las limitaciones del sistema desarrollado. Se concluye con recomendaciones específicas para trabajo futuro que permitan evolucionar esta prueba de concepto hacia un sistema de despliegue clínico.

Este trabajo abordó el desafío de desarrollar un sistema de asistencia diagnóstica en cáncer de mama que priorice interpretabilidad y actualizabilidad. Mediante la integración de Retrieval-Augmented Generation visual y textual con arquitecturas de orquestación multi-agente, se buscó demostrar que sistemas basados en evidencia recuperable pueden complementar modelos end-to-end tradicionales en entornos clínicos que requieren transparencia.

5.1 Implementación de Sistema Multiagente para Asistir Análisis De Imagen de Tejido WSI (Whole Slide Image)

A través de la implementación de un modelo generador y otro evaluador, apoyados con herramientas como RAG textual, clasificador ImageRAG e historial de conversación, se

creó un sistema multiagente capaz de apoyar la revisión de WSI por parte del patólogo. Lamentablemente, la precisión de estos modelos no es la adecuada para una tarea crítica como la enfrentada. Los modelos si bien son capaces de entregar respuestas coherentes y bien estructuradas, su conocimiento base no es suficiente para asistir en preguntas reflexivas que impliquen un conocimiento profundo del tema. Sin embargo, es necesario destacar que en el caso en donde el RAG textual encuentra información útil, como también cuando las preguntas son generales, el desempeño de los modelos tienen una precisión cercana al 100%.

El sistema diseñado implementa diversas herramientas como corrección de parches, visualización de parches y recuperación de imágenes similares, lo cual hace que la solución propuesta sea adecuada en un ambiente en donde la interpretabilidad es un requisito indispensable.

5.2 Recaudar Documentos e Imágenes Para Construcción de Sistema Retrieval Augmented Generation(RAG)

La recaudación de imágenes fue para el ImageRAG logrado completamente, implementando un RAG visual con imágenes de alta calidad, entregando un sólido ground truth para la tarea de clasificación. Lamentablemente, no se logró encontrar imágenes etiquetadas correspondientes al documento del College American Of Pathologist, que es una guía utilizada mundialmente.

Respecto a la recaudación de texto, se lograron encontrar diversas fuentes especializadas en Cáncer de mama en el contexto de análisis de WSI, sin embargo, no fue suficiente para responder las preguntas seleccionadas en nuestro apartado de pruebas. Esto nos indica que es necesario realizar un estudio extenso de las fuentes disponibles.

Nuestro clasificador ImageRAG logró resultados que igualan el estado del arte logrado en el concurso realizado por los creadores del dataset ICIAR 2018, aunque nuestra

solución incorpora interpretabilidad al usuario (apartado que estos trabajos no presentaron).

5.3 Evaluación de Distintos Modelos SML (Small Language Model).

A través de la selección de tres modelos SML, logramos evaluar la precisión y latencia. La elección de estos modelos especializados en el contexto médico aportaron positivamente a respuestas coherentes con el usuario objetivo (patólogos), sin embargo, debido a su cantidad de parámetros y entrenamiento base, no fueron capaces de entregar respuestas precisas en situaciones que requerían una reflexión intensa. Si contamos con un cuerpo extenso de texto de calidad en nuestro RAG textual, se podría lograr un desempeño superior.

5.4 Limitaciones

Las principales limitaciones enfrentadas corresponden al hardware existente y a la dificultad para conseguir información relacionada al análisis de WSI en cáncer de mama. Con un hardware moderno lograríamos disminuir los tiempos de ejecución, una recuperación más extensa ya sea de imágenes o texto, y permite implementar modelos más exigentes computacionalmente.

Respecto a la información existente, debido a que los datos médicos son sensibles, y crear un dataset etiquetado de calidad conlleva un tiempo considerable, existe una escasez de información de datos para el objetivo propuesto. Por lo mismo, no se fue capaz de armar un dataset propio que mejore la precisión obtenida. Gracias a la incorporación de la estrategia Human-In-The-Loop, eventualmente los patólogos crearán una base de datos de calidad que permita mejorar los datos existentes.

Finalmente, la ausencia de un patólogo como colaborador activo no permitió validar

la solución en un entorno realista. De igual manera, nuestra solución fue analizada por patólogos en su etapa final, en donde entregaron su feedback positivo a la solución diseñada.

5.5 Trabajo Futuro

Para expandir el sistema construido, se recomienda analizar técnicas para optimizar la latencia de nuestra solución, sin afectar el rendimiento logrado. Es necesario analizar los modelos elegidos en profundidad y realizar cambios a nivel de arquitectura si fuese necesario. Además, un análisis en detalle de hiperparámetros y de las estrategias implementadas, lograrían mejorar la robustez de los modelos.

Un estudio del estado del arte de sistemas multi-agentes permitiría aplicar soluciones más complejas, pero que mejoran la precisión de las respuestas obtenidas. El estudio de prompts, y de arquitectura de agentes, deberían contribuir a una mejora sustancial de nuestra solución.

Finalmente, es necesario recaudar una mayor cantidad de documentos para nuestro RAG textual, ya que queda demostrado que un corpus de conocimiento extenso proporciona a las respuestas contexto fundamental para responder. Es prioridad que los modelos tengan como contexto información validada por expertos, de forma que puedan citar la información recuperada. Esto entrega al sistema interpretabilidad visual y textual.

Bibliography

- [1] Elton Abels, Liron Pantanowitz, Famke Aeffner, Mark D. Zarella, Jeroen van der Laak, Marilyn M. Bui, Venkata N. Vemuri, Anil V. Parwani, Jeffrey Gibbs, Emilio Agosto-Arroyo, Andrew H. Beck, and Clive Kozlowski. Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association. *The Journal of Pathology*, 249(3):286–294, 2019.
- [2] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78, 2019.
- [3] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, Gerardo Fernandez, Jack Zeineh, Matthias Kohl, Christoph Walz, Florian Ludwig, Stefan Braunewell, Maximilian Baust, Quoc Dang Vu, Minh Nguyen Nhat To, Eal Kim, Jin Tae Kwak, Sameh Galal, Veronica Sanchez-Freire, Nadia Brancati, Maria Frucci, Daniel Riccio, Yaqi Wang, Lingling Sun, Kaiqiang Ma, Jiannan Fang, Ismael Kone, Lahsen Boulmane, Aurélio Campilho, Catarina Eloy, António Polónia, and Paulo Aguiar. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, 2019.
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan,

- Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [5] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024.
- [6] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pages 403–410, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- [8] Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, et al. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*, 2024.
- [9] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter,

- Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
- [10] Andrew J. Evans, Mohamed E. Salama, William H. Henricks, and Liron Pantanowitz. Implementation of whole slide imaging for clinical purposes: Issues to consider from the perspective of early adopters. *Archives of Pathology & Laboratory Medicine*, 141(7):944–959, 2017.
- [11] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.
- [12] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), October 2021.
- [13] Rim Guetari, Hana Ayari, and Hela Sakly. Computer-aided diagnosis systems: a comparative study of classical machine learning versus deep learning-based approaches. *Knowledge and Information Systems*, 65:3881–3921, 2023.
- [14] Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*, 2025.
- [15] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.
- [16] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.
- [17] Matti Katajamaki. Knowledge-based cad. *Expert Systems with Applications*, 3(2):277–287, 1991.

- [18] Asifullah Khan, Zunaira Rauf, Anabia Sohail, Amirr Rehman, Hifsa Asif, Aqsa Asif, and Umair Farooq. A survey of the vision transformers and their CNN-transformer based variants. *Artificial Intelligence Review*, 56(Suppl 3):2917–2970, 2023.
- [19] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48, 2020.
- [20] Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, et al. Small language models learn enhanced reasoning skills from medical textbooks. *NPJ digital medicine*, 8(1):240, 2025.
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [22] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
- [23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- [24] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

- [25] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Kenji Ikamura, Georg Gerber, Ivy Liang, Long Phi Le, Tong Ding, Anil V Parwani, et al. A foundational multimodal vision language ai assistant for human pathology. *arXiv preprint arXiv:2312.07814*, 2023.
- [26] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, 2009.
- [27] Gina Merino-Pereira. Puesta al día en cáncer de mama en Chile. *Revista chilena de obstetricia y ginecología*, 88:261 – 262, 10 2023.
- [28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- [29] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Medflamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.
- [30] Henning Müller, Andreas Holzinger, Markus Plass, Luka Brcic, Christoph Stumpner, and Kurt Zatloukal. Explainability and causability for artificial intelligence-supported medical image analysis in the context of the European In Vitro Diagnostic Regulation. *New Biotechnology*, 70:67–72, 2022.
- [31] R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt. Computer-aided detection of clustered microcalcifications on digital mammograms. *Medical & Biological Engineering & Computing*, 33(2):174–178, 1995.
- [32] Johnathan Pocock, Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Srijay Deshpande, Giorgos Hadjigeorgiou, Adam Shephard, Raja Muhammad Saad

- Bashir, Mohsin Bilal, Wenqi Lu, David Epstein, Fayyaz Minhas, Nasir M. Rajpoot, and Shan E. Ahmed Raza. TIAToolbox as an end-to-end library for advanced tissue image analytics. *Communications Medicine*, 2(1):120, October 2022.
- [33] C. Quinn, A. Maguire, and E. Rakha. Pitfalls in breast pathology. *Histopathology*, 82(1):140–161, 2023.
- [34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [35] Maral Rasoolijaberi, Morteza Babaie, Abtin Riasatian, Sobhan Hemati, Parsa Ashrafi, Ricardo Gonzalez, and Hamid Tizhoosh. Multi-magnification image search in digital pathology. *IEEE journal of biomedical and health informatics*, PP, 06 2022.
- [36] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [37] Konstantinos I Roumeliotis, Ranjan Sapkota, Manoj Karkee, and Nikolaos D Tselikas. Agentic ai with orchestrator-agent trust: A modular visual classification framework with trust-aware orchestration and rag-based reasoning. *arXiv preprint arXiv:2507.10571*, 2025.
- [38] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.
- [39] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

- [40] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.
- [41] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundström. Measuring domain shift for deep learning in histopathology. *IEEE Journal of Biomedical and Health Informatics*, 25(2):325–336, 2021.
- [42] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature Medicine*, 29:1930–1940, 2023.
- [43] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.
- [44] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.
- [45] Yawen Wu, Ming Cheng, Shuning Huang, Zixiao Pei, Yifeng Zuo, Jingxin Liu, Kai Yang, Qi Zhu, Jiawen Zhang, Huating Hong, Daoqiang Zhang, Kun Huang, Liang Cheng, and Wei Shao. Recent advances of deep learning for computational histopathology: Principles and applications. *Cancers*, 14(5):1199, 2022.
- [46] Julie Xiang, Xiusi Wang, Xuhai Zhang, Ying Xi, Fatoumata Eweje, Yiwen Chen, Yusheng Li, Christopher Bergstrom, Michael Gopaulchan, Taemin Kim, Kun-Hsing Yu, Sigal Willens, Francisco M. Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Serena Yang, and Ruijiang Li. A vision-language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

- [47] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Mu Wei Huang, Matthew P. Lungren, Nishant Naik, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024.
- [48] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.
- [49] Kai Zhang, Rong Zhou, Ershad Adhikarla, Zhiling Yan, Ming Liu, Jun Kuang, Chun-Nan Chen, Xianwen Li, Sirui Chen, Can Cui, Mingyu Wang, Jiayin Chen, Yuyin Liu, Yong Chen, Yupeng Chen, Xiaofan Yang, Fenglin Shen, Namu Zheng, Qingyu Huang, Xin Chen, Ruoqi Hong, Akshay S. Chaudhari, Alan L. Yuille, Zongwei Zhou, Yingda Wang, Yizhou Xie, David A. Clifton, Sheng Li, Chenyu You, and Kang Xu. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30:3129–3141, 2024.
- [50] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.
- [51] Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, pages 4442–4457, 2025.

Apéndice

Apéndice A: Textos utilizados para poblar la base vectorial

El corpus de conocimiento para el sistema RAG textual fue construido utilizando los siguientes textos:

1. National Comprehensive Cancer Network (2023). *NCCN Clinical Practice Guidelines in Oncology: Breast Cancer*.
2. International Agency for Research on Cancer (2019). *WHO Classification of Tumours: Breast Tumours*, 5th Edition.
3. American Cancer Society (2024-2025). *Breast Cancer Facts & Figures*.
4. Rosen, P.P. (2014). *Rosen's Breast Pathology*, 4th Edition.
5. Colomer et al. *Biomarkers in breast cancer: A consensus statement by the Spanish Society of Medical Oncology and the Spanish Society of Pathology*.
6. DeVita, Hellman, and Rosenberg's. *Cancer, Principles & Practice of Oncology*.
7. European Society of Radiology. *Breast Imaging*.
8. Cancer Care Alberta (2021). *Follow-Up Care for Early-Stage Breast Cancer*.
9. Pan American Health Organization, World Health Organization. *Prevention: Breast Cancer Risk Factors And Prevention*.
10. Susan G. Komen *What Is Breast Cancer?*.

Apéndice B: Whole Slide Image utilizada para prueba de latencia según tamaño de parches



Figure 5.1: Whole Slide Image utilizada para prueba de latencia en clasificación. Esta muestra fue proporcionada por Environ SPA.

Apéndice C: Prompt utilizado para evaluar la respuesta de la arquitectura de agentes

El siguiente prompt corresponde a un ejemplo utilizado para evaluar la respuesta de la arquitectura de agentes:

```
1 I need you to evaluate the response of an LLM "evaluator" (who takes  
   the generator's response and improves it). For the evaluation, use  
   the following rubric:
```

Relevance: To what extent does the response directly address the problem or question?

Accuracy: The degree to which the response provides correct, reliable, and fact-based information?

Appropriateness: How well does the response suit the target audience while being respectful in tone and content?

6 Completeness: The degree to which the response covers the essential aspects of the question?

Coherence: How logically structured and well-organized is the response ?

Clarity: How easily can the response be understood by the target audience?

Please rate my answers to this rubric between 0 and 5. Finally, I want to know if incorporating the evaluator model improves the initial response.

11

Question:

I'm analyzing a cribriform lesion that looks like low-grade DCIS.

However, the 'lumens' are filled with a fibrillar, slightly basophilic material rather than secretion.

16

If these 'spaces' are actually enclosed by myoepithelial cells rather than luminal epithelial cells, what is the specific diagnosis, and what material makes up the 'spherules' ?

Correct Answer (Ground Truth):

21

Diagnosis : Collagenous Spherulosis .

Material : The spherules are composed of Basement Membrane material (Collagen IV) , not mucin. 3 Distinction : In DCIS , the cribriform spaces are true lumens lined by

epithelial cells . In Collagenous Spherulosis , the material is
26 extracellular matrix produced by myoepithelial cells .

Initial Answer:

The Answer is: The lesion described is likely a solid papillary
carcinoma in situ (SPCIS). The 'lu-
31 mens' filled with fibrillar, slightly basophilic material are
characteristic of the solid component of SPCIS. The diagnosis is
based on the presence of atypical epithelial cells forming solid
clusters or sheets without an intact myoepithelial layer, as seen
in conventional papillary DCIS. The term 'spherules' is not
typically used in this context.

Final Answer:

The user is asking for a diagnosis based on a description of a
cribriform lesion with specific features:
36 - Resembles low-grade DCIS.
- 'Lumens' filled with fibrillar, slightly basophilic material (not
secretion).
- The spaces are enclosed by myoepithelial cells, not luminal
epithelial cells.

The draft response suggests Solid Papillary Carcinoma In Situ (SPCIS).
Let's evaluate this and consider other possibilities based on the
provided information and general knowledge.

- 41
1. ****DCIS (Ductal Carcinoma In Situ):**** The user mentions it looks like low-grade DCIS, but the filling material is not typical secretion. DCIS is characterized by malignant epithelial cells confined to the ducts. Different grades (low, intermediate, high) and patterns (cribriform, solid, papillary, micropapillary) exist.
 2. ****Papillary DCIS:**** This involves malignant cells forming papillary structures within the ducts. It can be low, intermediate,

- or high grade.
3. ****Solid Papillary Carcinoma In Situ (SPCIS):**** This is a subtype of papillary DCIS characterized by solid nests of cells, often lacking a true lumen or myoepithelial layer. It typically involves intermediate or high-grade cells. The description of "fibrillar, slightly basophilic material" filling the spaces could potentially fit the cellular material seen in SPCIS, although it's not a classic description.
 4. ****Myoepithelial Cells:**** The key information is that the spaces are enclosed by *myoepithelial cells*. This is crucial. In typical DCIS (including papillary DCIS), the malignant cells are luminal epithelial cells, and the myoepithelial layer is usually intact (though it can be discontinuous, especially in higher grades or invasive disease). If the spaces are lined by myoepithelial cells, it suggests the malignant cells are *outside* the myoepithelial layer, which would imply invasion. However, the user describes a "lesion" that *looks like* low-grade DCIS, suggesting it might be *in situ*.
 - 46 5. ****Invasive Carcinoma:**** If the malignant cells are outside the myoepithelial layer, it's invasive carcinoma. The pattern described (cribriform-like spaces lined by myoepithelial cells) could represent certain types of invasive carcinoma, such as invasive ductal carcinoma with cribriform features or tubular carcinoma.
 6. ****Dimorphic Papillary DCIS:**** This is a specific subtype where both epithelial and myoepithelial cells form the papillary structures. The cells often resemble myoepithelial cells but have epithelial features. This might fit the description if the "fibrillar, slightly basophilic material" represents the mixed cell population.
 7. ****Other possibilities:**** Could it be a benign lesion mimicking DCIS? Less likely given the description.

Let's re-evaluate the draft's suggestion of SPCIS. SPCIS typically involves solid nests of *epithelial* cells, often lacking a true lumen and myoepithelial layer. The description mentions "spaces"

lined by myoepithelial cells. This contradicts the typical SPCIS morphology where the malignant cells *are* the solid nests, and the myoepithelial layer is absent or discontinuous.

51

The description "*spaces... enclosed by myoepithelial cells*" is the most critical piece of information. This strongly suggests the malignant cells are *outside* the myoepithelial layer. If the lesion is confined (in situ), this is unusual. If it's invasive, it points towards invasive carcinoma.

However, the user states it *looks like* low-grade DCIS. This might imply the overall architecture is ductal or cribriform, but the cellular details are different.

56

Let's reconsider the provided RAG snippets:

- **Snippet 1 (DCIS grades):** Describes different grades and patterns of DCIS. Mentions microcalcifications and correlation with radiology. Doesn't directly address the specific morphology described.
- **Snippet 2 (DCIS and invasion):** Mentions distinguishing DCIS from lobular carcinoma in situ and invasive carcinoma. Highlights the importance of myoepithelial cell markers to confirm intact layer in DCIS. States absence of reactivity doesn't unequivocally distinguish DCIS from invasive carcinoma. Mentions irregular cell clusters suggest invasion.
- **Snippet 3 (Papillary DCIS):** Discusses papillary DCIS, including dimorphic papillary DCIS where both epithelial and myoepithelial cells are involved. Mentions immunohistochemistry for distinguishing subtypes.

61

The description "*spaces... enclosed by myoepithelial cells*" is highly suggestive of invasive carcinoma, where the malignant cells have breached the myoepithelial layer. The "*fibrillar, slightly basophilic material*" could be the malignant cells themselves or associated stromal components. The term "*spherules*" is not standard

terminology here.

Given the description:

- Cribriform-like pattern.
- Spaces lined by myoepithelial cells (implying malignant cells are outside this layer).
- 66 - Fibrillar, basophilic material filling the spaces.
- Resembles low-grade DCIS overall.

This combination points strongly towards an invasive carcinoma, possibly with a cribriform or tubular pattern, where the malignant cells are outside the myoepithelial layer. The "fibrillar, slightly basophilic material" could represent the malignant cells themselves or a combination of cells and matrix.

- 71 The draft's answer (SPCIS) doesn't fit well with the "spaces enclosed by myoepithelial cells" description. SPCIS typically lacks a myoepithelial layer surrounding the solid nests of malignant cells.

Therefore, a more accurate answer should address the implication of myoepithelial cells lining the spaces.

Revised Answer Construction:

- 76
1. Acknowledge the user's description.
 2. Highlight the significance of myoepithelial cells lining the spaces.
 3. Explain that this feature typically indicates invasion, as malignant cells are outside the myoepithelial layer.
 4. Suggest possible diagnoses based on the cribriform pattern and material: Invasive Ductal Carcinoma (cribriform pattern) or Tubular Carcinoma.
 - 81 5. Address the "spherules" term - it's not standard; perhaps refers to cell clusters or calcifications.
 6. Mention the need for further evaluation (immunohistochemistry, etc



.) to confirm the diagnosis and grade.

****Final Answer Formulation:****

86 The description of spaces lined by myoepithelial cells, rather than luminal epithelial cells, is a key finding. In ductal carcinoma in situ (DCIS), the malignant cells are confined within the ductal system, typically surrounded by an intact myoepithelial layer. When spaces are lined by myoepithelial cells, it usually indicates that the malignant cells have breached this layer and are therefore invasive.

The lesion described, with a cribriform pattern and fibrillar, basophilic material filling the spaces lined by myoepithelial cells, is most consistent with an ****invasive carcinoma****. Possible specific diagnoses include:

1. ****Invasive Ductal Carcinoma (IDC)** with

91

Try to make the answer as concise and complete as possible, highlighting only the most important points.