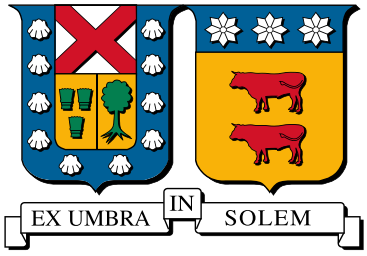
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE ELECTRÓNICA

VALPARAÍSO - CHILE



"DESARROLLO DE UNA INTERFAZ DE VISUALIZACIÓN DE IMÁGENES MÉDICAS PARA EL DIAGNÓSTICO EN ONCOLOGÍA DE PRECISIÓN."

CLAUDIO FRANCISCO ANDRÉS ZANETTA PENNA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRÓNICO.

PROFESOR GUIA: PhD. WERNER CREIXELL FUENTES

PROFESOR CORREFERENTE: PhD. 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: Desarrollo de una interfaz de visualización de imágenes médicas para el diagnóstico en oncología de precisión.

Nombre del candidato(a): Claudio Francisco Andrés Zanetta Penna

Carrera / Grado: Ingeniero Civil Electrónico

Campus: Casa Central **Departamento:** Departamento de Electrónica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Werner Greixell, 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: 27/04/2026 **Firma:** Greixell

Estudiante o Candidato(a):

Fecha: 13 de abril de 2026 **Firma:** C Zanetta

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.

Para Isabella y Carmen.

Agradecimientos

Es difícil condensar en palabras lo agradecido que estoy con todas las personas que han sido participes, de una forma u otra, en mi vida. En un intento de mantenerlo breve, me gustaría comenzar agradeciendo a mi abuelita, Carmen Bugueño, por todos los años que cuidó de mi y a quién nunca dejaré de extrañar.

Me gustaría agradecer a mi madre, Pierina Penna, y a mi padre, Jorge Zanetta. Ambos sacrificaron muchas cosas en sus vidas para que no me faltase nada. Si soy lo que soy, se los debo a ellos.

Me gustaría agradecer a mis hermanas, Catalina Zanetta e Isabella Zanetta, por apoyarme en todas las decisiones que he tomado.

A mis amigos, *los tomo*, por todas las risas y por estar tanto en los buenos momentos como en los malos.

A Josefina Vera y Tomás Riveros, por brindarme su gran amistad, decisiva en estos últimos años.

A mis profesores y maestros, por depositar su confianza en mi persona.

Y tanto a mis nuevas amistades como a las que he perdido a lo largo de estos años.

Cada uno de ustedes me ha brindado algún que otro pequeño momento de felicidad, y es la suma de estos lo que me hace seguir en pie, muchas gracias.

DEPARTAMENTO DE ELECTRONICA
UNIVERSIDAD TECNICA FEDERICO SANTA MARIA

DEPARTAMENTO DE ELECTRONICA

Desarrollo de una interfaz de visualización de imágenes médicas para el diagnóstico en oncología de precisión

Claudio Francisco Andrés Zanetta Penna

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

Universidad Técnica Federico Santa María

Profesor Guía: PhD. Werner Creixell Fuentes

Profesor Correferente: PhD. Alejandro Weinstein Oppenheimer

MARZO 2026

Resumen

El biomarcador Ki-67 es un indicador relevante en la prognosis del cáncer de mama; sin embargo, el conteo y cálculo de su índice de proliferación constituye un proceso costoso y demandante. En este trabajo se propone una interfaz y dos *pipelines* complementarios: uno orientado a la generación de nuevos *datasets* a partir de anotaciones nucleares mediante la extracción de parches positivos, negativos y de fondo, y otro destinado a la inferencia, realizando el conteo de instancias Ki-67 positivas y negativas y la estimación del índice de proliferación. La interfaz gráfica permite visualizar explícitamente las segmentaciones y clasificaciones realizadas, aportando explicabilidad visual al usuario. La interfaz funciona por defecto utilizando Cellpose SAM y un clasificador ConvNeXt-Tiny entrenado sobre el *dataset* público BC-Data, alcanzando un *F1-score* global de 84.17 % y un RMSE de 0.08076. La interfaz permite el cambio de segmentador y clasificador según sea necesario. En conjunto, el sistema demuestra la utilidad de un *MVP* extensible y adaptable a otros biomarcadores o tareas de patología digital.

Palabras Clave: Visión por Computador, Prognosis, Cáncer de mama, Segmentación, Interfaz

Development of a medical image visualization interface for precision oncology diagnosis

Claudio Francisco Andrés Zanetta Penna

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

Universidad Técnica Federico Santa María

Thesis Advisor: PhD. Werner Creixell Fuentes

Co-Advisor: PhD. Alejandro Weinstein Oppenheimer

MARCH 2026

Abstract

The Ki-67 biomarker is a relevant indicator in breast cancer prognosis; however, the counting and calculation of the proliferation index constitute a costly and time-consuming process. In this work, a graphical interface and two complementary *pipelines* are proposed: one oriented toward the generation of new *datasets* from nuclear annotations through the extraction of positive, negative, and background patches, and another focused on inference, performing the counting of Ki-67 positive and negative instances and estimating the corresponding proliferation index. The graphical interface enables explicit visualization of the performed segmentations and classifications, providing visual explainability to the user. By default, the interface operates using Cellpose SAM as the segmenter and a ConvNeXt-Tiny classifier trained on the public BCData *dataset*, achieving a global *F1-score* of 84.17 % and an RMSE of 0.08076. The interface allows the segmenter and classifier to be replaced as required. Overall, the system demonstrates the usefulness of an extensible *MVP* adaptable to other biomarkers or digital pathology tasks.

Keywords: Computer Vision, Prognosis, Breast Cancer, Segmentation, Interface.

Glosario

CNN Convolutional Neural Network. Tipo de red neuronal diseñada para el procesamiento de imágenes mediante operaciones convolucionales jerárquicas.

CPU Central Processing Unit. Procesador de propósito general encargado de la ejecución de tareas secuenciales y de control del sistema.

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

H&E Hematoxilina y Eosina. Técnica de tinción histológica estándar donde la hematoxilina resalta núcleos celulares y la eosina tiñe citoplasma y tejido conectivo.

H-DAB Hematoxilina–DAB. Modalidad de tinción inmunohistoquímica que combina hematoxilina con DAB para visualizar la expresión de biomarcadores como Ki-67.

IHQ Inmunohistoquímica. Técnica basada en anticuerpos para detectar proteínas específicas en tejidos, utilizada en la evaluación de biomarcadores tumorales.

Ki-67 Proteína nuclear asociada a proliferación celular, utilizada como biomarcador pronóstico en cáncer de mama.

Patch Parche, subimagen extraída alrededor de una región de interés, utilizada como entrada para modelos, en este caso, de clasificación celular.

Precision Precisión. Métrica que mide la proporción de verdaderos positivos respecto al total de predicciones positivas del modelo.

Recall Sensibilidad. Métrica que cuantifica la proporción de verdaderos positivos detectados respecto al total de positivos reales.

RMSE Root Mean Square Error. Raíz del error cuadrático medio, utilizada para medir la discrepancia entre valores predichos y valores de referencia.

Segmentación Proceso de identificación y delimitación de regiones de interés dentro de una imagen, como núcleos celulares.

VRAM Video Random Access Memory. Memoria dedicada de la GPU utilizada para almacenar modelos, tensores y activaciones durante el cómputo.

WSI Whole Slide Image. Imagen digital de alta resolución de un tejido histológico completo, escaneada a magnificaciones típicas de 20× o 40×.

Dataset Conjunto estructurado de datos utilizado para entrenamiento, validación o evaluación de modelos, que incluye imágenes y, cuando corresponde, anotaciones asociadas.

Ground truth Anotaciones de referencia consideradas correctas, utilizadas como base para el entrenamiento y la evaluación del desempeño del modelo.

Pipeline Secuencia organizada de etapas de procesamiento que transforma los datos o elementos de entrada en resultados finales.

Macro F1-score
Promedio no ponderado del F1-score calculado por clase, otorgando igual importancia a cada clase independientemente de su frecuencia.

Micro F1-score
F1-score calculado a partir del total global de verdaderos positivos, falsos positivos y falsos negativos, reflejando el desempeño general del modelo en conjuntos desbalanceados.

PyTorch Biblioteca de deep learning de código abierto utilizada para la implementación y entrenamiento de modelos de aprendizaje profundo, basada en grafos computacionales dinámicos y con soporte nativo para GPU.

Índice de contenidos

Lista de Figuras

Lista de Tablas

1 Introducción

1.1 Contexto y Motivación

En histopatología, el análisis de biopsias y muestras tisulares requiere la aplicación de técnicas de tinción que permitan diferenciar las estructuras del tejido. La tinción con Hematoxilina y Eosina (H&E) es la más utilizada, ya que permite identificar con claridad la morfología celular y tisular. En este contexto, la inmunohistoquímica constituye una técnica complementaria orientada a detectar la presencia de marcadores proteicos o antígenos específicos en un tejido tumoral.

En el caso del cáncer, uno de los biomarcadores más relevantes para la prognosis y la definición del tratamiento es el Ki-67, cual registra la velocidad a la cual se dividen las células cancerosas, ampliamente utilizado, entre otros, en el cáncer de mama. Un enfoque habitual para su detección es la técnica H-DAB, donde el cromógeno 3,3'-diaminobencidina tiñe de color marrón los núcleos que expresan el biomarcador, mientras que los núcleos negativos se observan en tonos azules. A partir de esta tinción, es posible calcular el índice de proliferación Ki-67, el cual, en conjunto con otros exámenes, influye de manera significativa en la decisión terapéutica y en la intensidad del tratamiento a aplicar.

La correcta estimación de este índice es crítica, ya que sesgos en el conteo o en la clasificación de núcleos positivos y negativos pueden afectar directamente la interpretación clínica. Sin embargo, el proceso de anotación manual de células positivas y negativas resulta exhaustivo, costoso en tiempo y no está exento de errores, incluso cuando es realizado por expertos.

Tradicionalmente, este análisis se realizaba mediante observación directa al microscopio. En la actualidad, la digitalización de las muestras histológicas en forma de láminas completas (*Whole Slide Images*, WSI) ha permitido avances significativos en el análisis computacional de tejidos. Estas imágenes, que pueden alcanzar resoluciones del orden de gigapíxeles a magnificaciones de 20x o 40x, han abierto nuevas posibilidades para el análisis automatizado y el apoyo al diagnóstico.

En este escenario, la visión por computador aplicada a patología digital surge como una herramienta prometedora para asistir en la detección y cuantificación de biomarcadores. Sin

embargo, persisten dificultades tanto para el patólogo como para los sistemas automáticos, asociadas a la variabilidad entre muestras, diferencias en protocolos de tinción, condiciones de adquisición de las imágenes, heterogeneidad del tejido y superposición celular.

Motivado por estas problemáticas, este trabajo propone un pipeline modular basado en segmentación y clasificación de núcleos, orientado a la cuantificación automática del índice Ki-67, específicamente para el cáncer de mama. El enfoque emplea herramientas de código abierto y modelos de *Deep learning*, con el objetivo de evaluar su desempeño en distintos conjuntos de datos y explorar su integración en una interfaz interactiva como prototipo funcional. El alcance del trabajo es de carácter experimental y académico, orientado a analizar fortalezas, limitaciones y proyecciones futuras del sistema propuesto, sin pretensión de reemplazar la evaluación realizada por un patólogo experto.

1.2 Problemática y desafíos

La problemática actual, específicamente en el contexto de este trabajo y de la visión por computador aplicada a la estimación del índice de proliferación Ki-67, se centra en dos ejes principales.

1.2.1 Escasez de *Datasets* públicos.

En primer lugar, existe una escasez de conjuntos de datos públicos y la ausencia de estándares unificados respecto a los procesos de tinción, escaneo, resolución y formato de almacenamiento. Esto se debe principalmente a que la información de los pacientes es privada y, por tanto, los datos quedan restringidos a hospitales o laboratorios que realizan los análisis. Las características de las imágenes generadas dependen del equipamiento utilizado, de las decisiones de compresión, de la magnificación, del tamaño de los parches extraídos e incluso de variaciones en los protocolos de tinción.

Bajo estas condiciones, tampoco es posible conocer con certeza el nivel de calidad de las anotaciones disponibles. Estas dependen del criterio del patólogo y, aunque se asumen correctas, no están exentas de error o preferencias. Dado el carácter privado de la mayoría de estos conjuntos de datos, no existe un mecanismo externo de verificación o validación cruzada.

Si bien existen iniciativas que liberan datasets públicos, estos presentan una alta heterogeneidad entre sí. No existe un estándar común, lo que dificulta que un modelo entrenado en un único conjunto de datos generalice adecuadamente a otros. Además, las anotaciones varían significativamente: algunos datasets entregan coordenadas de núcleos, otros solo conteos globales, y otros proporcionan parches asociados a un porcentaje de Ki-67 sin indicar la localización exacta de las células. Esta diversidad de formatos complica el desarrollo de modelos generales y comparables entre trabajos.

1.2.2 Falta de explicabilidad visual.

Una segunda problemática, estrechamente ligada a la anterior, es que desde el punto de vista clínico el objetivo final es la estimación confiable del índice de proliferación Ki-67, independientemente del método utilizado para obtenerlo. Sin embargo, enfoques basados únicamente en la predicción directa del índice o en el análisis de proporciones de canales de color carecen de explicabilidad visual, obligando al usuario a confiar en el modelo como una caja negra.

El enfoque basado en la segmentación de núcleos y su posterior clasificación surge como una alternativa que permite otorgar explicabilidad visual al patólogo, facilitando la comprensión de cómo se obtuvo un determinado valor del índice. No obstante, este enfoque requiere anotaciones más detalladas, al menos a nivel de centro celular o regiones nucleares, lo que incrementa significativamente el costo y el esfuerzo de generación de datos, reforzando nuevamente el problema de escasez de datasets adecuados.

1.3 Solución Propuesta

Este trabajo propone un pipeline modular y configurable, diseñado para adaptarse a distintos conjuntos de datos y niveles de calidad de anotación. El enfoque permite reemplazar o ajustar sus componentes en función de los datos disponibles, manteniendo un desempeño comparable, aunque no superior, al estado del arte. En su configuración base, el sistema utiliza Cellpose SAM como segmentador y ConvNeXt-Tiny como clasificador.

1.3.1 *Pipeline* de modificacion de datasets

Se propone un pipeline de modificación de datasets orientado a la generación de parches de entrenamiento. A partir de imágenes con anotaciones de células positivas y negativas, se emplea un segmentador generalista para extraer máscaras nucleares y generar parches correspondientes a las clases positivo, negativo y fondo. Esta separación resulta clave para entrenar un clasificador capaz de discriminar correctamente entre células relevantes y detecciones no informativas, sin restringir el sistema a un único biomarcador.

1.3.2 *Pipeline* modular de segmentación y detección

El pipeline de segmentación y clasificación permite procesar nuevas imágenes de forma modular. Las máscaras generadas por el segmentador son clasificadas como positivas, negativas o fondo, y a partir de las clases relevantes se estima el índice de proliferación Ki-67 por imagen. Esta estructura permite sustituir el clasificador por modelos más robustos o especializados, así como extender el enfoque a otros biomarcadores.

1.3.3 Interfaz de visualización con anotaciones al usuario

Se implementa una interfaz de visualización desarrollada en Python utilizando PySide6, que integra el pipeline de inferencia completo. La aplicación permite visualizar las máscaras segmentadas, las clasificaciones positivas y negativas, y el índice Ki-67 resultante, entregando un nivel de explicabilidad visual que facilita la interpretación de los resultados por parte del usuario.

1.4 Objetivos

1. **Desarrollar un sistema a partir de modelos de *Deep Learning*, *Computer Vision* y técnicas afines para la detección de biomarcadores en imágenes de inmunohistioquímica IHQ.**

2. **Implementar técnicas de explicabilidad visual para validar los resultados de los modelos entrenados.**

3. **Diseñar una interfaz gráfica para cargar las imágenes a evaluar con el fin de poder procesarlas y generar una imagen de salida con anotaciones al usuario.**

1.5 Organización del Documento

El resto de este documento se estructura como sigue:

- **Capítulo 2 - Estado del Arte:** Revisión histórica de la visión por computador aplicada en medicina, patología digital, prognosis en tejido mamario y trabajos relacionados sobre la estimación del índice de proliferación d Ki-67.
- **Capítulo 3 - Diseño e Implementación de la Solución:** Pipeline propuesto para entrenamiento e inferencia, minería de datos, pruebas para determinar las configuraciones y módulos óptimos del pipeline y desarrollo de la interfaz.
- **Capítulo 4 - Resultados y Análisis:** Análisis adicionales que justifican la elección del modelo de clasificación y comparación de desempeño con otros trabajos, análisis de conjuntos de datos y limitaciones del prototipo.
- **Capítulo 5 - Conclusiones y Trabajo Futuro:** Presentación de las conclusiones del trabajo, evaluación del cumplimiento de los objetivos, limitaciones del sistema, proyección de mejoras y desarrollo futuro.

2 Estado del Arte

El presente capítulo revisa la historia y el estado actual del desarrollo de la visión por computador aplicada a la medicina. Se abordan los fundamentos teóricos que constituyen la base conceptual del presente trabajo. Asimismo, se revisa el estado actual de la detección de biomarcadores, con énfasis en el marcador Ki-67 y su aplicación en la prognosis del tejido mamario.

Adicionalmente, se analizan investigaciones relevantes en el área, las cuales aportan antecedentes que permiten contextualizar el problema y justificar las decisiones metodológicas adoptadas.

Finalmente, se presentan definiciones y resúmenes breves de los distintos enfoques considerados, junto con un análisis de sus principales fortalezas y limitaciones.

2.1 Visión por Computador y uso en medicina

La visión por computador (*Computer Vision*) es una rama de la inteligencia artificial cuyo objetivo es permitir que los sistemas computacionales analicen e interpreten información visual. Para ello, emplea técnicas de procesamiento de imágenes y métodos de aprendizaje automático. Estas técnicas permiten abordar tareas que van desde la detección de bordes hasta la localización, detección y segmentación de instancias.

2.1.1 Visión por computador y evolución

Los orígenes de la visión por computador se sitúan alrededor de la década de 1960. Uno de los primeros hitos fue el estudio de la inferencia de estructuras tridimensionales a partir de imágenes bidimensionales. Este enfoque quedó evidenciado en el trabajo pionero de Lawrence Roberts, quien demostró el potencial de los sistemas computacionales para extraer información estructural desde imágenes [51].

En 1980, Kunihiko Fukushima propuso una nueva arquitectura en su trabajo *Neocognitron*. En este modelo se establecieron las bases conceptuales de lo que posteriormente se conocería como redes neuronales convolucionales [20].

Con la introducción de LeNet a comienzos de la década de 1990, Yann LeCun presentó una de las primeras redes neuronales convolucionales funcionales [39]. En el contexto de este trabajo también se introdujo el conjunto de datos MNIST, el cual se consolidó como un *benchmark* ampliamente utilizado en múltiples publicaciones del área [71].

A pesar de los avances propuestos por LeCun, el uso de redes neuronales convolucionales no se consolidó hasta las décadas siguientes. Durante este periodo, las tareas de visión se enfocaron en el desarrollo de descriptores y extractores de característica así como algoritmos de clasificación de aprendizaje automático. Entre estos se encuentran las *Support Vector Machines* (SVM), el método de *k-Nearest Neighbors* (k-NN), los procesos gaussianos, entre otros [1, 5].

En 2009 se presentó el conjunto de datos ImageNet [14]. Este conjunto de datos, junto con el aumento en la capacidad de cómputo disponible en esa época, permitió el entrenamiento de modelos de mayor complejidad.

En 2012 ocurrió otro hito relevante con la publicación del trabajo AlexNet. Este modelo popularizó el uso de GPU para el entrenamiento y destacó por su arquitectura basada en capas convolucionales y funciones de activación ReLU [35].

AlexNet consolidó el uso de redes neuronales profundas, en particular redes neuronales convolucionales, para tareas de visión por computador.

En los años posteriores se introdujeron nuevas técnicas y familias de modelos, como los transformadores de visión y arquitecturas híbridas.

2.1.2 Diagnóstico asistido por computador

Los primeros avances de la visión por computador en medicina se centraron en sistemas de diagnóstico asistido por computador (*Computer-Aided Diagnosis*, CAD). Estos enfoques se basaban principalmente en técnicas de procesamiento de imágenes.

Dichas técnicas se aplicaron inicialmente a estudios radiológicos. Entre ellos se incluyen radiografías, tomografías computarizadas y resonancias magnéticas [16, 22].

En una etapa posterior y siguiendo el avance de la visión por computador general, los sistemas comenzaron a incorporar algoritmos de aprendizaje automático para mejorar la detección y clasificación de patrones clínicamente relevantes. En particular, métodos como las *Support Vector Machines* (SVM) fueron ampliamente utilizados en combinación con descriptores de características para tareas de clasificación y segmentación en imágenes médicas [37].

Tras la introducción de arquitecturas profundas y el éxito de modelos como AlexNet, el uso de redes neuronales profundas en imágenes médicas aumentó significativamente [9].

Desde entonces, las redes neuronales convolucionales se consolidaron como una herramienta para el análisis de imágenes médicas. Un ejemplo representativo es la arquitectura U-Net [52]. Su diseño tuvo como objetivo principal la segmentación de imágenes médicas.

Con los años diversos otros enfoques han coexistido, orientados a trasformadores de visión u otras arquitecturas, con el objetivo de perfeccionar tareas de detección, segmentación y cuantificación de estructuras anatómicas y patológicas.

2.2 Arquitecturas y paradigmas en segmentación y detección de imágenes

Si bien los modelos propuestos buscan resolver tareas similares, sus supuestos y mecanismos de aprendizaje difieren técnicamente, así como su capacidad para capturar información visual.

En el contexto de la visión por computador aplicada a la medicina, las redes neuronales convolucionales han sido ampliamente utilizadas en tareas de segmentación y detección. No obstante, el surgimiento reciente de los transformadores de visión ha introducido enfoques alternativos que han sido evaluados en aplicaciones médicas [33].

2.2.1 Redes Convolucionales

Las redes neuronales convolucionales son una alternativa preponderante en el ámbito médico. Su adopción se debe al desempeño alcanzado por diversas arquitecturas propuestas en la literatura. Entre ellas, se destacan las siguientes:

2.2.1.1 VGG : Introducida en 2014 por Simonyan et al. [55], esta arquitectura planteó como cambio relevante el uso sistemático de filtros de convolución de 3×3. Este diseño permitió construir redes más profundas con un uso más eficiente de los parámetros.

Si bien su uso ha disminuido con el paso de los años en favor de otras arquitecturas [15], VGG aún se emplea en distintos trabajos como parte de pipelines híbridos. Un ejemplo de ello es el trabajo de Chen et al. [11], donde se propone una fusión entre VGG-16 y modelos de transformadores de visión para la clasificación de tumores óseos a partir de imágenes de tomografía computarizada.

2.2.1.2 U-Net Presentada en 2015 por Ronneberger et al., esta arquitectura propone un esquema *encoder–decoder*. El diseño incorpora etapas de *down-sampling* y *up-sampling*,Este enfoque permite la segmentación de células y tejidos dentro de una imagen [52].

A pesar de su antigüedad, U-Net se mantiene relevante, esto se evidencia en trabajos como nnU-Net [29] y D2HU-Net [65]. En ellos, un mejor ajuste de parámetros y modificaciones arquitectónicas permiten alcanzar resultados cercanos al estado del arte.

2.2.1.3 ResNet Introducida en 2015 por He et al. [24], en Microsoft Research, esta arquitectura propone el uso de conexiones residuales. Estas conexiones permiten que la salida de una etapa sea una función de la transformación aprendida sumada a la entrada original, mediante una *shortcut connection*. Este diseño facilita el entrenamiento de redes profundas y mitiga problemas como la atenuación o explosión del gradiente.

Existen variantes de esta arquitectura, como ResNeXt [70], la cual introduce el concepto de cardinalidad como un nuevo eje de diseño. Dicha extensión busca mejorar la capacidad representacional sin aumentar significativamente la complejidad del modelo.

ResNet continúa siendo utilizada en aplicaciones médicas. Un ejemplo es el trabajo de Das et al., donde el uso de ResNet-50 alcanzó una precisión de clasificación del 92.01% en la diferenciación entre tejido mamario sano y maligno en imágenes de resonancia magnética [13].

2.2.1.4 DenseNet Presentada en 2016 por Huang et al. [26], esta arquitectura introduce conexiones densas entre sus capas. En este esquema, la salida de cada capa se utiliza como entrada para todas las capas posteriores.

Este diseño favorece la reutilización de características y una propagación más estable de la información a lo largo de la red.

DenseNet continúa siendo utilizada en la literatura actual. Un ejemplo es el trabajo de Md. Alamin Talukder, donde se emplea DenseNet169 para la detección de cáncer de mama [59].

2.2.1.5 EfficientNet Desarrollada por Tan et al., esta arquitectura introduce el concepto de *compound scaling*. Este enfoque propone escalar de manera conjunta la profundidad, el ancho y la resolución del modelo. El objetivo es maximizar el rendimiento manteniendo un equilibrio entre precisión y complejidad computacional [60].

El trabajo original presenta siete modelos preentrenados con distintos números de parámetros. Posteriormente, se propuso una actualización denominada EfficientNetV2 [61].

El uso de EfficientNet se ha extendido al ámbito médico. Un ejemplo es el modelo MoEffNet [3], donde EfficientNet se emplea como extractor de características. En dicho trabajo se reporta un F1-score superior al 99.0% para la detección de cáncer de mama en mamografías.

Otro ejemplo es el trabajo de Latha et al., donde mediante un ajuste fino de EfficientNet-B7 se alcanza una precisión de clasificación del 99.14% en imágenes ultrasónicas de tejido mamario [38].

2.2.2 Transformadores de Visión

La arquitectura de los transformadores se originó en el procesamiento de lenguaje natural [64]. Su incorporación al ámbito de la visión por computador se produjo con el trabajo de Dosovitskiy et al. [17].

En este enfoque, en lugar de modelar relaciones entre *tokens* lingüísticos, se utilizan parches de una imagen. Cada parche es proyectado a un espacio de *embeddings* y tratado como un *token* visual.

Sobre esta representación se aplica el mecanismo de autoatención para modelar relaciones globales entre distintas regiones de la imagen.

Uno de los primeros desarrollos relevantes basados en *Vision Transformer* fue DeiT (*Data-efficient Image Transformers*). Este modelo fue propuesto como una extensión directa de *Vision Transformer*, orientada a reducir la dependencia de grandes volúmenes de datos y recursos computacionales.

Para ello, DeiT introduce un esquema de entrenamiento basado en *knowledge distillation* a partir de un modelo maestro [63].

Posteriormente, surgieron los denominados modelos fundacionales. Estos se caracterizan por su entrenamiento a gran escala y su capacidad de generalización a múltiples tareas.

Dentro de este contexto, se introdujeron modelos que incorporan información visual junto con el lenguaje.

En el ámbito de la visión por computador, enfoques como BEiT [8] y CLIP [50] ejemplifican esta tendencia. Estos modelos emplean transformadores entrenados sobre grandes volúmenes de datos. Esto permite aprender representaciones visuales reutilizables y transferibles a nuevas tareas.

Bajo esta idea, se desarrollaron diversos modelos fundacionales entrenados y ajustados fina-

mente con conjuntos de datos médicos. Entre ellos se encuentran MedCLIP [66], MUSK [68], MedGemma [53] y MedLlama [69].

En la actualidad, estos modelos no representan la única aplicación de los transformadores de visión. Un ejemplo adicional es Segment Anything Model (SAM) [34]. Este modelo, a partir de un encoder basado en Vision Transformer, permite la segmentación de distintos tipos de imágenes de forma general.

2.3 Procesamiento digital de imágenes patológicas

La Digital Pathology Association (DPA) define la patología computacional (*Computational Pathology*, CPATH) como un enfoque orientado al análisis de grandes volúmenes de datos en patología. En este contexto, múltiples fuentes de información de un paciente, incluyendo imágenes y metadatos, se combinan para extraer patrones y analizar características relevantes [2].

Para efectos prácticos, este trabajo se enfoca en una subárea de la patología computacional, correspondiente al análisis de *Whole Slide Images* (WSI).

2.3.1 Whole Slide Images

Las *Whole Slide Images* (WSI) corresponden a imágenes médicas de alta resolución obtenidas mediante escáneres digitales. Estas imágenes operan en el orden de los gigapíxeles.

Una imagen escaneada con una magnificación de 40× presenta típicamente un tamaño entre 1 y 4 GB [49]. Existen casos donde imágenes sin compresión, con magnificación 20× y una resolución de 0.220 μm/píxel, pueden alcanzar tamaños cercanos a los 10 gigapíxeles. En estos escenarios, el almacenamiento requerido puede llegar a aproximadamente 30 GB por imagen, dificultando su manejo computacional [7].

Una solución común es la compresión de las imágenes. Para ello, se emplean formatos sin pérdida, como JPEG2000, o con pérdida, como JPEG. Mediante estos esquemas, el tamaño de una WSI puede reducirse hasta aproximadamente 500 MB en el primer caso.

Otro enfoque ampliamente utilizado consiste en dividir la WSI en múltiples parches que se

analizan de forma independiente. Comúnmente se emplean parches de 256×256 píxeles [32]. Esta estrategia permite reducir el costo computacional asociado al procesamiento de estas imágenes. En la figura 2.1 se presenta una imagen que describe los niveles de magnificación de las *WSI* y sus usos comunes por nivel.

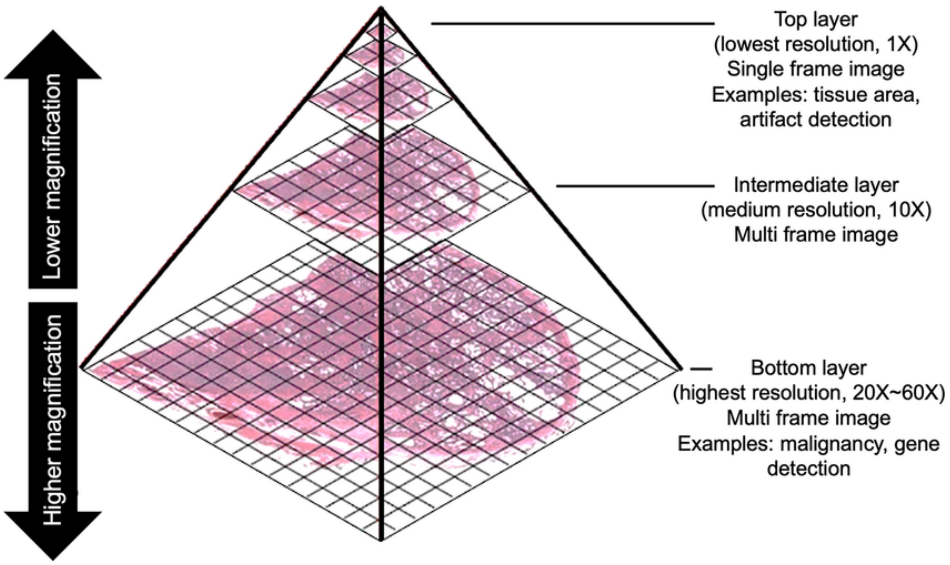


Figura 2.1: Modelo de pirámide multicapa para imágenes de Whole Slide Images (WSI). Fuente: Hossain et al. (2023) [25]. Imagen utilizada bajo licencia CC BY 4.0.

2.3.2 Tinción H&E y su uso en patología digital

La tinción de hematoxilina y eosina (H&E) constituye el método estándar para el análisis histopatológico de tejidos [46]. Esta técnica permite resaltar la morfología celular y tisular mediante el contraste entre núcleos y citoplasma. Los núcleos se tiñen en tonalidades violetas, mientras que el citoplasma adquiere tonalidades rosadas. En la Figura 2.2 se presenta un parche de 1000×1000 píxeles donde se evidencian las tonalidades violeta y rosadas características de la tinción H&E.

Tradicionalmente, el análisis de láminas teñidas con H&E se realizaba mediante microscopía óptica convencional [10]. En este contexto, la evaluación dependía de la observación directa del patólogo. El análisis implicaba recorrer manualmente distintas regiones del tejido para identificar patrones morfológicos relevantes.

En la actualidad, la digitalización de estas láminas histológicas ha dado lugar al uso de *Whole*

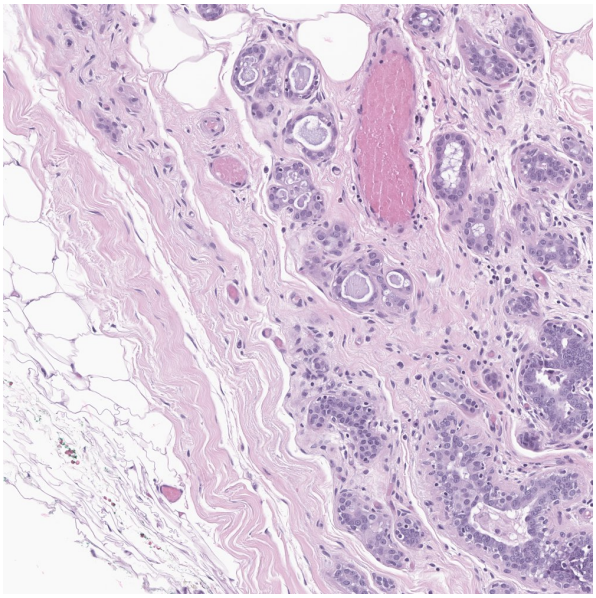


Figura 2.2: Parche de WSI de tejido mamario con tinción H&E. Fuente: *Dataset IHC4BC Compressed*, Akbarnejad et al. (2023) [4].

Slide Images (WSI). Este proceso introduce un compromiso entre fidelidad visual y manejabilidad computacional, debido a la discretización y compresión inherentes a la digitalización. No obstante, la representación digital de las muestras permite su almacenamiento, análisis remoto y reutilización.

Además, las WSI habilitan su uso como entrada para métodos de aprendizaje automático y aprendizaje profundo. Esto permite entrenar y evaluar modelos computacionales sobre grandes volúmenes de tejido histológico, abriendo nuevas posibilidades para el análisis automatizado.

2.3.3 Inmunohistioquímica

Por otro lado, la inmunohistoquímica (IHC) corresponde a una técnica que permite detectar antígenos en un tejido, comúnmente proteínas, mediante el uso de anticuerpos que se unen al antígeno de interés [30]. Esta técnica es ampliamente utilizada en el diagnóstico oncológico, tanto para fines de prognosis como para la detección de biomarcadores.

De acuerdo con el cromógeno utilizado en la reacción antígeno–anticuerpo, se genera una tinción visible. Un enfoque ampliamente empleado es el uso de 3,3′-diaminobenzidina (DAB). En este caso, una reacción positiva se visualiza en color marrón.

Para proporcionar contraste con las células sin reacción, se emplea hematoxilina como contratinción, dando origen a la tinción H-DAB. Esta combinación permite visualizar las estructuras negativas en tonalidades azules [19]. En la figura 2.3 se presenta un ejemplo de esta tinción en un parche de resolución 1000×1000, donde se observa la detección del biomarcador Ki-67.

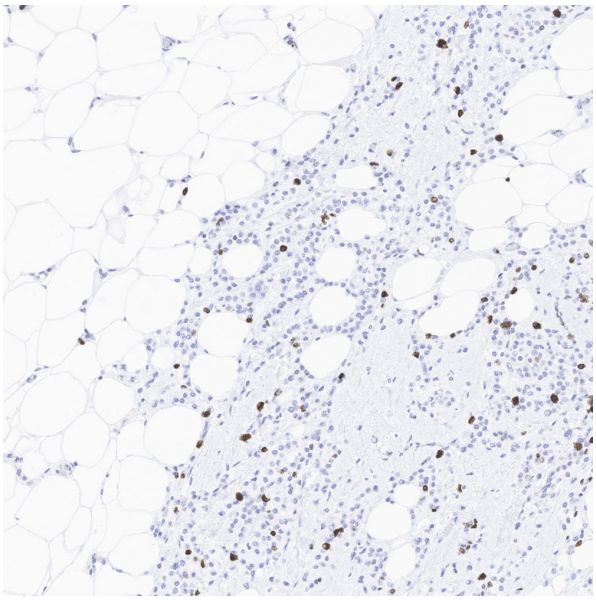


Figura 2.3: Parche de WSI de tejido mamario con tinción H-DAB. Fuente: *Dataset IHC4BC Compressed*, Akbarnejad et al. (2023) [4].

2.4 Prognosis en cáncer de tejido mamario

Existen diversos métodos para el diagnóstico y la prognosis del cáncer de mama. Sin embargo, este trabajo se enfoca en el uso de biomarcadores relevantes para este fin.

2.4.1 Biomarcadores relevantes y diagnóstico

Según el Instituto Nacional del Cáncer de Estados Unidos [47], los biomarcadores permiten entregar información sobre la agresividad del cáncer. En particular, pueden indicar qué tan rápido puede crecer un tumor, qué tan probable es que se propague a otras partes del cuerpo y qué tan efectiva puede ser una terapia determinada.

Entre los biomarcadores más utilizados en la práctica clínica se encuentran:

2.4.1.1 ER & PR: Estos biomarcadores indican si el crecimiento tumoral es estimulado por las hormonas estrógeno (ER) y progesterona (PR). Los tumores con receptores hormonales positivos (HR+) pueden tratarse mediante terapias hormonales que bloquean estas señales de crecimiento. En contraste, los tumores HR- no responden a este tipo de terapias y suelen requerir quimioterapia. El estado de estos receptores puede cambiar con el tiempo, lo que puede hacer necesarias nuevas biopsias.

2.4.1.2 HER2: La proteína HER2 participa en el control del crecimiento celular. Su sobreexpresión (HER2+) se asocia a una división celular acelerada y a un comportamiento tumoral más agresivo. La determinación de este estado es fundamental para identificar la elegibilidad del paciente a terapias dirigidas contra esta proteína.

2.4.1.3 Ki-67: Ki-67 es un biomarcador de proliferación celular. Se expresa exclusivamente en células que se encuentran en proceso de división. Un índice Ki-67 elevado indica una alta tasa proliferativa. Este valor se utiliza como un indicador clínico relevante de agresividad tumoral y de posible sensibilidad a la quimioterapia.

2.5 Estado actual del uso de visión por computador para apoyo diagnóstico en cáncer de mama mediante imágenes IHC con el biomarcador Ki-67

El presente trabajo no tiene como objetivo analizar los distintos tipos de tratamiento. Su enfoque se centra en la estimación de métricas asociadas al estado de los biomarcadores. Asimismo, se busca aportar mecanismos de explicabilidad visual que permitan justificar los resultados obtenidos.

Entre los distintos biomarcadores disponibles, se seleccionó Ki-67. Esta elección se fundamenta en la existencia de trabajos previos y conjuntos de datos relevantes en la literatura. Además, Ki-67 no solo es útil para la prognosis del cáncer de mama, sino que corresponde a un biomarcador transversal a otros tipos de cáncer [40].

2.6 Trabajos relacionados

En la actualidad, dado que para el diagnóstico suele ser suficiente estimar el índice Ki-67 y clasificarlo como alto, medio o bajo, numerosos trabajos optan por calcular el porcentaje de células positivas y negativas en una imagen. Este enfoque permite obtener una medida global del índice de proliferación. Sin embargo, la definición de un valor alto o bajo de Ki-67 puede variar según el protocolo clínico y el centro de análisis. En algunos casos, valores superiores al 30% son considerados indicativos de alta proliferación [28].

Además, este tipo de estimación carece de explicabilidad a nivel de instancia, ya que el resultado depende de la predicción global del modelo sin una validación directa célula a célula.

Trabajos como el de Kukučka et al. [36] proponen esquemas de entrenamiento débilmente supervisados. En este enfoque, el modelo se entrena utilizando una imagen junto con un valor global de proliferación. El método permite identificar regiones relevantes y estimar el índice Ki-67. No obstante, la explicabilidad visual se limita a mapas de calor, sin incluir conteo ni marcado explícito de instancias.

Otro enfoque consiste en la estimación del índice Ki-67 a partir de imágenes teñidas con H&E. Un ejemplo es el trabajo de Liu et al. [43], donde se reporta una precisión de 0.9371 en la discriminación entre células positivas, negativas y fondo.

No obstante, existen trabajos orientados al conteo explícito de instancias celulares. Entre ellos se encuentra PiNet [21], que mediante el uso de conjuntos de datos privados y mejoras en las anotaciones de conjuntos de datos públicos, junto con el diseño de una red convolucional, logra una precisión del 86% en la estimación del índice de proliferación de Ki-67.

Asimismo, trabajos como PathoNet [48], basados en una arquitectura U-Net, permiten no solo estimar el índice Ki-67, sino también calcular el índice de proliferación de linfocitos infiltrantes del tumor (TIL). Por otra parte, el trabajo de Anglada-Rotger et al. [6] propone un enfoque de conteo celular que utiliza HoverNet [23] para la generación del *ground truth*, seguido de segmentación y clasificación mediante un modelo basado en dos redes U-Net.

2.6.1 Modelos, conjuntos de datos y limitaciones actuales

Diversos trabajos proponen arquitecturas y conjuntos de datos distintos. Sin embargo, no todos los modelos ni los datos utilizados son públicos o de libre acceso.

Este es el caso de PiNet y del trabajo de Anglada-Rotger et al. En ambos estudios se describe la arquitectura empleada, pero no se liberan los modelos entrenados ni los conjuntos de datos utilizados durante el entrenamiento. En particular, aunque PiNet utiliza conjuntos de datos públicos como DeepSlides [54], el equipo realizó una anotación propia del estado positivo o negativo de cada núcleo.

En contraste, PathoNet propone y libera un conjunto de datos con anotaciones explícitas. Dicho conjunto incluye las coordenadas de los centroides de cada núcleo, lo que permite el entrenamiento y evaluación de modelos orientados al conteo celular.

Cabe volver a destacar que no existe una estandarización en los conjuntos de datos disponibles. Esto se debe, en primer lugar, a la variabilidad introducida por los distintos tipos de escáneres utilizados para la adquisición de WSI [18]. Además, existen diferencias asociadas a los protocolos de inmunohistoquímica empleados para la tinción de los tejidos [30, 41]. Como consecuencia, los conjuntos de datos pueden variar significativamente entre distintos centros de investigación.

Por otro lado, existen conjuntos de datos cuyas anotaciones no permiten una explicabilidad visual a nivel de instancia. Un ejemplo es el conjunto de datos IHC4BC [4]. Si bien este conjunto proporciona el número total de células en una imagen y el nivel de tinción asociado a cada una, no especifica las coordenadas de los núcleos. Esto se debe a que su objetivo es la predicción del índice Ki-67 a partir de imágenes H&E, lo cual resulta suficiente para un análisis global del índice de proliferación.

En este contexto, es común encontrar conjuntos de datos con anotaciones parciales o diseñadas para un modelo o línea de trabajo específica.

Existen dos conjuntos de datos que presentan anotaciones adecuadas para el presente trabajo. El primero corresponde a SHIDC-B-Ki-67, propuesto junto con PathoNet. El segundo es BCData, presentado por Huang et al. [27]. Si bien estos conjuntos no incluyen máscaras

celulares completas, las anotaciones de centroides han resultado suficientes en trabajos previos, cabe destacar que estos conjuntos de datos tienen anotaciones a nivel de parche y no de WSI.

Finalmente, el trabajo de Liu et al. [42] propone un modelo orientado a mejorar la explicabilidad visual y el conteo de instancias. Este enfoque se entrena utilizando los conjuntos de datos SHIDC-B-Ki-67 y BCData. No obstante, el modelo entrenado no se encuentra disponible públicamente, lo que limita su reutilización.

Con base en estas consideraciones, el presente trabajo toma como referencia los enfoques y conclusiones de la literatura previa para el diseño y evaluación de la propuesta metodológica.

2.6.2 Contribuciones de este trabajo

El presente trabajo propone un *pipeline* modular y de fácil modificación para el análisis de biomarcadores en imágenes histopatológicas. El diseño por módulos permite sustituir el modelo utilizado según el biomarcador de interés y el conjunto de datos disponible.

El sistema se entrena utilizando conjuntos de datos públicos y de libre acceso. A partir de una imagen de entrada, el modelo permite detectar biomarcadores positivos y negativos, realizar el conteo de instancias celulares y estimar el índice de proliferación Ki-67. Los resultados obtenidos son comparables con el estado del arte reportado en la literatura.

Adicionalmente, se integra un prototipo de interfaz gráfica de uso sencillo, compatible con sistemas operativos Windows y Linux.

Las principales contribuciones de este trabajo se resumen a continuación:

- Propuesta de un *pipeline* modular para la detección y cuantificación de biomarcadores en imágenes histopatológicas.
- Propuesta de *pipeline* de adaptación de conjuntos de datos públicos y de libre acceso para el entrenamiento y evaluación de los modelos.
- Implementación de un enfoque basado en conteo explícito de instancias celulares para la estimación del índice Ki-67.

- Incorporación de mecanismos de explicabilidad visual mediante la identificación y clasificación de células positivas y negativas.
- Desarrollo de un prototipo de interfaz gráfica multiplataforma para facilitar el uso del sistema.

3 Diseño e Implementación de la Solución

En este trabajo se proponen dos *pipelines* diferenciados. El primero está orientado a la construcción y adaptación de conjuntos de datos. El segundo corresponde al *pipeline* de inferencia utilizado en la interfaz final.

El *pipeline* de construcción de conjuntos de datos permite adaptar otros *datasets* para entrenar modelos, siempre que estos cumplan con las anotaciones requeridas. Este enfoque facilita el entrenamiento de nuevos modelos ante la disponibilidad de nuevos datos o biomarcadores.

Tal como se presenta en la Figura 3.4, este *pipeline* se basa en un modelo de segmentación. A partir de centros celulares anotados como positivos y negativos, el segmentador permite generar máscaras celulares. Estas máscaras se utilizan posteriormente para extraer parches individuales. Dichos parches constituyen el conjunto de entrenamiento para un modelo de

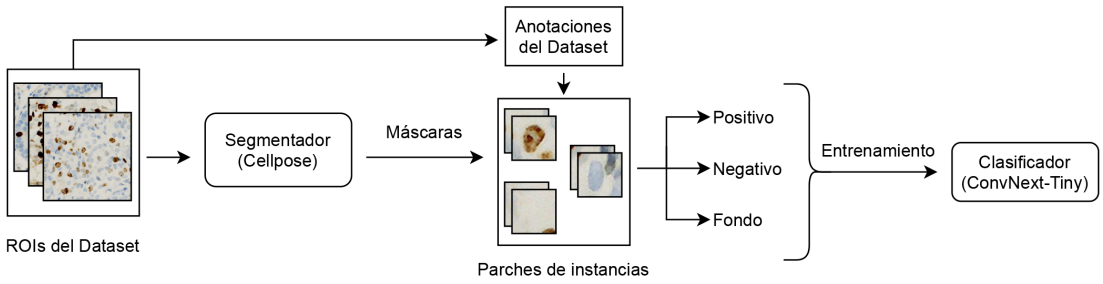


Figura 3.4: *Pipeline* para la adaptación de *datasets* y entrenamiento del modelo clasificador. Parches e imágenes de muestra provenientes del *Dataset* de BCData [27].

clasificación de células positivas, negativas y fondo.

El segundo *pipeline* está orientado a la inferencia y al uso final del sistema. Este es el flujo empleado en la interfaz de usuario. Como se muestra en la Figura 3.5, el proceso comienza con un segmentador maestro. Este modelo segmenta las células presentes en la imagen y genera parches individuales. Posteriormente, un clasificador discrimina cada parche y asigna la clase correspondiente.

La motivación para este diseño, surge la necesidad de contar con la mayor cantidad de ejemplos posibles para cada clase, con los cual entrenar un modelo que pueda discernir entre estas. Luego, el usar nuevamente un segmentador general en la parte de inferencia, permite que el

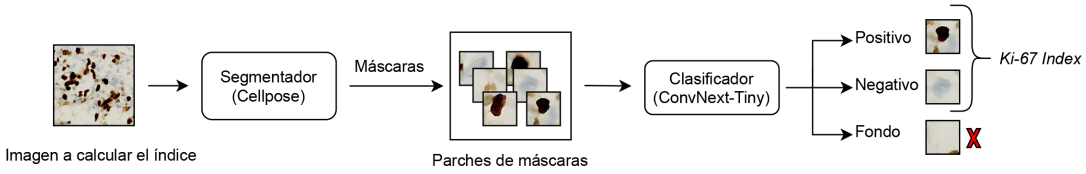


Figura 3.5: *Pipeline* de inferencia *segmentador y clasificador*, empleado en la interfaz. Parches e imágenes de muestra provenientes del *Dataset* de BCData [27].

modelo entrenado pueda dedicarse solamente a clasificar y discriminar que es relevante y que no.

La idea de estructurar la solución como el *pipeline* descrito, no surgió como primera opción. Durante la exploración de distintas alternativas para el desarrollo de la interfaz y la posibilidad de ampliar su uso, y a partir del análisis de enfoques propuestos en trabajos relacionados, se concluyó que esta estrategia resultaba válida. No obstante, desde un inicio se consideró necesaria una etapa de adaptación de los datos que permitiera generar *datasets* más completos.

En principio, disponer de datos completos, como los obtenidos en la etapa de adaptación de *dataset*, permitiría entrenar un único modelo que realice todas estas tareas de manera conjunta. Este es el enfoque seguido por modelos como PathoNet.

Sin embargo, PathoNet presenta limitaciones prácticas para su integración en este trabajo. El código se encuentra desactualizado y opera sobre versiones antiguas de Python. Además, el modelo realiza todas las tareas dentro de una arquitectura monolítica. Esto dificulta la actualización o sustitución de componentes sin reentrenar el sistema completo.

La separación del flujo en dos etapas independientes permite resolver estas limitaciones. Por un lado, una etapa de segmentación y detección. Por otro, una etapa de clasificación. Este diseño modular facilita la sustitución progresiva de modelos a medida que se dispone de mejores arquitecturas o datos de entrenamiento.

Este enfoque resulta especialmente relevante en el contexto médico. Los datos clínicos suelen ser privados y están sujetos a restricciones de acceso. En muchos casos, los centros de investigación pueden trabajar con estos datos bajo la condición de mantenerlos confidenciales. El *pipeline* propuesto permite entrenar nuevos modelos de forma eficiente cuando se dispone de datos de calidad, sin necesidad de modificar el flujo completo ni la interfaz.

La separación de tareas también se observa en otros trabajos. Un ejemplo es HoVerNet, propuesto por Graham et al. [23]. En este modelo, las tareas de segmentación, detección y clasificación se realizan de forma paralela dentro de un único sistema. HoVerNet se ha consolidado como una referencia en la literatura al abordar el problema de células adyacentes detectadas como una sola instancia, trabajos previos solucionaban este desafío mediante algoritmos como *watershed*, como es el caso de Pathonet, o *stardist* que ha mostrado buen desempeño en esta tarea como indica el trabajo de Weigher et al. [67], e incluso se emplea en el tarbajo de Akbarnejad et al. [4] para la segmentación del *nuclei* en la confección de su *dataset*. Trabajos posteriores orientados a HoVerNet, como los de Zhang et al. [72] y Tang et al. [62], demuestran que el modelo puede mejorar su desempeño mediante ajustes y optimizaciones adicionales.

Inicialmente, se consideró el uso de HoVerNet como solución integral. No obstante, este enfoque presenta dificultades para el ajuste fino en el contexto de este trabajo. HoVerNet requiere anotaciones específicas y se encuentra entrenado principalmente con imágenes H&E. En consecuencia, no captura adecuadamente la señal cromogénica asociada a tinciones DAB, lo que limita su aplicación directa en imágenes IHC. Entrenar el modelo desde cero implicaría perder el conocimiento previamente aprendido, por lo que esta alternativa fue descartada.

Enfoques alternativos utilizan HoVerNet como herramienta para la construcción de conjuntos de datos. Un ejemplo es el trabajo de Anglada-Rotger et al.[6], donde HoVerNet se emplea para segmentar células en un conjunto de datos privado, que posteriormente se utiliza para entrenar un modelo basado en U-Net dual. Aunque el uso directo de HoVerNet no se ajusta a los objetivos de este trabajo, se adopta la idea de emplear un modelo de segmentación para generar conjuntos de datos mejor anotados.

Siguiendo este razonamiento, y basándose en el trabajo de Liu et al., se consolidó la idea del primer *pipeline*, encargado de la generación parches celulares que permitan entrenar un modelo enfocado exclusivamente en detección y clasificación. Para ello, basta con contar con anotaciones de positividad, negatividad y fondo, lo que simplifica el proceso de generación de datos y permite separarlo respecto a la tarea de clasificación.

Para la etapa de segmentación se seleccionó Cellpose [57]. Este modelo de aprendizaje pro-

fundo permite segmentar células de distintos tipos sin requerir entrenamiento específico por tejido. Esta característica resulta especialmente útil para generar conjuntos de datos más completos . Además, este enfoque permite incorporar explícitamente la clase de fondo, ausente en muchos conjuntos de datos existentes. Esto se debe a que Cellpose segmenta células de manera independiente del tipo de imagen o tinción. La decisión de considerar o descartar estas máscaras queda delegada al clasificador, el cual determina si el parche corresponde a una célula positiva, negativa o a fondo. Este diseño contribuye a una mayor robustez del sistema.

En febrero de 2025 se publicó Cellpose 3 [58]. Posteriormente, en diciembre de 2025, se presentó Cellpose-SAM [45], el cual integra Segment Anything Model para una segmentación más robusta. Ambas variantes fueron evaluadas y comparadas para determinar su uso en la construcción de los conjuntos de datos y en la etapa de segmentación de la interfaz.

Bajo este contexto, en las siguientes secciones se describe el proceso de modificación de los conjuntos de datos SHIDC-B-Ki-67 y BCData. Estas modificaciones permiten entrenar modelos de detección celular y evaluar su desempeño. Finalmente, se presenta la comparación de los modelos obtenidos y la selección del más adecuado para su integración en la interfaz final.

Para la etapa de desarrollo se utilizaron las métricas de *recall*, *precision*, *accuracy* y *F1-score* con el fin de comparar el desempeño de los distintos modelos y arquitecturas evaluadas. La elección de cada métrica depende del tipo de tarea abordada (clasificación, detección o segmentación) y de la disponibilidad de anotaciones completas en el *ground truth*. Estas métricas se definen a partir de los siguientes conceptos fundamentales:

- **Verdadero Positivo (TP)**: corresponde a una célula correctamente detectada y clasificada como perteneciente a la clase positiva o negativa, según corresponda.
- **Falso Positivo (FP)**: corresponde a una detección clasificada como célula relevante cuando, según la verdad de terreno, no corresponde a una célula de interés o pertenece a otra clase.
- **Falso Negativo (FN)**: corresponde a una célula presente en la verdad de terreno que no fue detectada o fue clasificada incorrectamente.
- **Verdadero Negativo (TN)**: corresponde a una región correctamente identificada como

fondo, es decir, una instancia que no contiene una célula relevante y fue clasificada correctamente como tal.

A partir de estas definiciones, las métricas utilizadas se calculan de la siguiente forma:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.1}$$

El *recall* mide la capacidad del modelo para recuperar las instancias relevantes presentes en el conjunto de datos. Un valor alto de *recall* indica que el modelo omite pocas células verdaderas, lo cual resulta especialmente importante en el contexto de la cuantificación de biomarcadores, donde la omisión de células positivas puede conducir a una subestimación del índice de proliferación.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

La *precision* mide la proporción de detecciones positivas que son correctas. Esta métrica resulta particularmente relevante en tareas de detección y segmentación, donde no se dispone de una anotación exhaustiva del fondo y, por tanto, los verdaderos negativos no están explícitamente definidos. En este contexto, la *precision* permite cuantificar el nivel de ruido introducido por detecciones incorrectas, penalizando la presencia de falsos positivos.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

La *accuracy* representa la proporción total de predicciones correctas realizadas por el modelo. Esta métrica resulta adecuada únicamente cuando el conjunto de datos está completamente etiquetado y existen verdaderos negativos bien definidos, como ocurre durante la etapa de entrenamiento del clasificador de parches, donde cada muestra pertenece explícitamente a una de las clases consideradas (fondo, Ki-67 positivo o Ki-67 negativo). Sin embargo, en tareas de detección o segmentación, donde únicamente se anotan las instancias presentes y no el fondo de manera exhaustiva, la *accuracy* deja de ser representativa y no se utiliza durante la

evaluación en inferencia.

$$\text{F1-score} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{3.4}$$

El *F1-score* corresponde a la media armónica entre *precision* y *recall*, y permite evaluar el equilibrio entre la capacidad del modelo para recuperar instancias relevantes y el control de errores de clasificación. Esta métrica resulta especialmente útil en escenarios con desbalance entre clases o cuando tanto los falsos positivos como los falsos negativos tienen un impacto significativo en la interpretación del resultado.

En este trabajo, si bien se reportan distintas métricas según la etapa del pipeline, se prioriza el *F1-score* sobre las clases Ki-67 positivo y Ki-67 negativo, dado que este se encuentra directamente relacionado con la estimación del índice de proliferación celular. Un balance adecuado entre *precision* y *recall* en estas clases resulta fundamental para evitar tanto la sobreestimación como la subestimación del biomarcador, asegurando una cuantificación más robusta y consistente.

3.1 Entorno de experimentación

Todos los experimentos desarrollados en este trabajo fueron ejecutados en una estación de trabajo local dedicada, utilizada tanto para el entrenamiento de modelos como para las etapas de validación, prueba y desarrollo de la interfaz. El sistema operativo empleado fue Linux Mint 19.1, basado en Ubuntu 18.04 LTS. El equipo cuenta con un procesador Intel Core i7-8700K de 6 núcleos físicos y 12 hilos, operando a una frecuencia base de 4.3 GHz, acompañado de 16 GB de memoria RAM. Para la aceleración de los modelos de aprendizaje profundo se utilizó una GPU NVIDIA basada en la arquitectura GV102, con soporte CUDA, empleando el driver NVIDIA versión 530.30.02. El sistema dispone de aproximadamente 2.26 TB de almacenamiento local.

El desarrollo experimental se apoyó en tres entornos virtuales independientes, gestionados mediante Miniconda y con el objetivo de aislar dependencias y evitar conflictos entre bibliotecas. El *framework* elegido para cada entorno fue *PyTorch*.

El primer entorno fue utilizado para la evaluación y ajuste de Cellpose 3, el segundo para las pruebas con Cellpose SAM, y el tercero para la ejecución del pipeline completo y de la interfaz gráfica desarrollada. Esta separación fue necesaria debido a diferencias en las dependencias requeridas por cada versión del segmentador, así como por el uso de bibliotecas adicionales asociadas a la interfaz y a la inferencia del modelo.

Los paquetes y versiones utilizados en cada entorno se encuentran especificados mediante archivos de configuración (`.yml`) disponibles en el repositorio asociado a este trabajo, lo que permite la reproducción completa del entorno experimental 5.2.

3.2 Modificación de Datasets

Cellpose fue seleccionado como segmentador principal debido a su integración directa mediante *pip*, su carácter generalista y su buen desempeño en distintos contextos. No obstante, dado que este trabajo considera múltiples conjuntos de datos con distintas características, fue necesario evaluar su desempeño en tareas de segmentación de manera independiente para cada uno de ellos.

El conjunto de datos SHIDC-B-Ki-67 presenta dos resoluciones distintas: una variante con parches de 256×256 píxeles y otra con parches de 1228×1228 píxeles, las cuales se denominan en este trabajo, por simplicidad, como variantes SD y HD, respectivamente. También según convenga, se anotarán por su nombre completo «SHIDC-B-Ki-67» o como «SHIDC». Por otro lado, el conjunto de datos BCData utiliza parches de 640×640 píxeles. En total, se consideran tres configuraciones de datos distintas. Cada una de ellas presenta no solo diferencias de resolución, sino también variaciones en el tamaño aparente de las células, producto de la magnificación utilizada durante el proceso de escaneo.

Cellpose permite adaptarse a estas diferencias siempre que se ajusten adecuadamente un conjunto de parámetros que influyen directamente en su desempeño. En este trabajo se consideraron los siguientes:

1. **Diámetro celular**: tamaño promedio esperado de las células e píxeles, utilizado por el algoritmo para escalar internamente el proceso de detección. Para SHIDC-B-Ki-67 SD

se utilizó un valor de 15, para SHIDC-B-Ki-67 HD un valor de 72 y para BCData un valor de 47.

2. **Diámetro mínimo**: cota inferior del tamaño de las detecciones, en píxeles, bajo la cual una máscara es descartada.

3. ***Cell probability threshold***: o abreviadamente "*prob*", controla cuán permisivo es el modelo al detectar células; valores menores incrementan el número de detecciones.

4. ***Flow threshold***: o abreviadamente "*flow*", regula la tolerancia del modelo frente a variaciones morfológicas en las máscaras detectadas; mayores valores incrementan la tolerancia.

Dado que en esta etapa resulta prioritario maximizar la cantidad de parches disponibles por conjunto de datos, con el fin de disponer de una mayor diversidad de ejemplos para el entrenamiento del clasificador, se realizó una comparación entre las dos versiones de Cellpose utilizadas en este trabajo: Cellpose 3 y Cellpose-SAM. Esta comparación incluyó análisis tanto cualitativos como cuantitativos para los tres conjuntos de datos considerados.

El procedimiento seguido constó de tres etapas. En primer lugar, se realizó una selección del modelo base más adecuado. En segundo lugar, se ajustaron los parámetros de segmentación mediante un análisis cualitativo, evaluando visualmente los resultados sobre una imagen representativa del conjunto de entrenamiento. Finalmente, se llevó a cabo una evaluación cuantitativa utilizando el subconjunto de entrenamiento completo, analizando cuántas células anotadas en el conjunto de verdad de terreno coincidían con las máscaras generadas por el modelo. Los dos primeros análisis fueron de carácter cualitativo, mientras que el último fue cuantitativo.

3.2.1 Cellpose 3

Cellpose 3 dispone de tres modelos principales entrenados para distintos tipos de estructuras: *cyto3*, *tissuenet* y *nuclei*. Para seleccionar el modelo más adecuado, se realizó una comparación cualitativa de su desempeño sobre los distintos conjuntos de datos. A partir de este análisis se observó que el modelo *cyto3* era capaz de segmentar una mayor cantidad de células sin

necesidad de ajustes finos de parámetros, por lo que fue seleccionado como modelo base. Un ejemplo representativo de esta comparación se muestra en la Figura 3.6, donde *cyto3* logra detectar un total de 272 células.

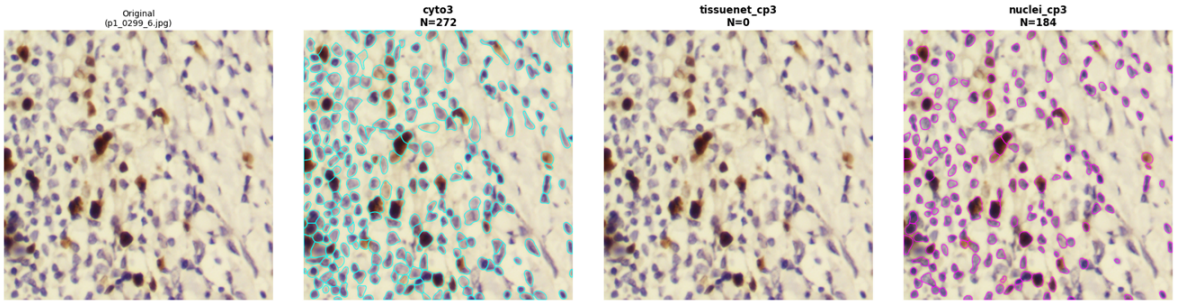
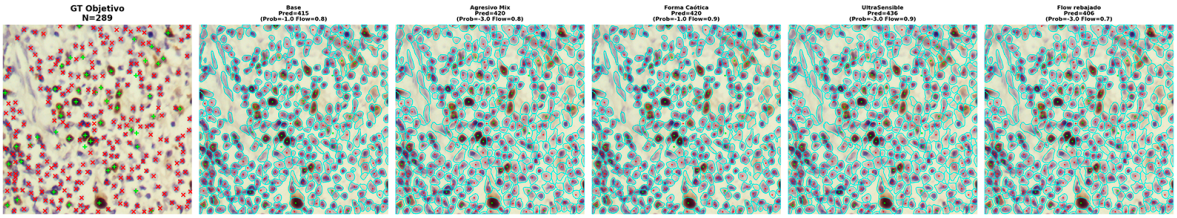
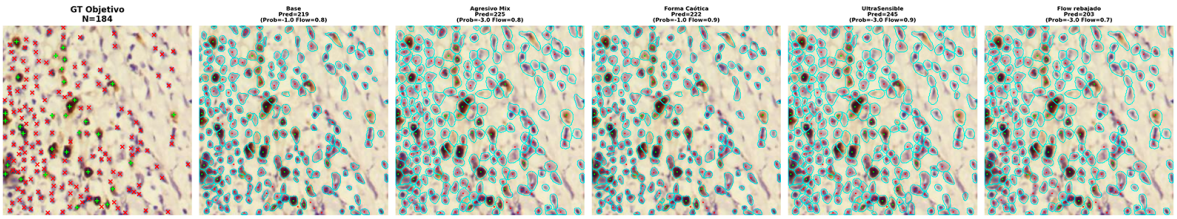


Figura 3.6: Comparación de los 3 modelos principales de cellpose-3 en una imágen de prueba.

Una vez seleccionado el modelo *cyto3*, se evaluaron distintas combinaciones de parámetros para cada conjunto de datos. Este análisis se realizó de forma cualitativa, considerando el equilibrio entre la recuperación de células y la introducción de ruido. Las Figuras 3.7 ilustran este proceso para las variantes HD y SD de SHIDC-B-Ki-67, mientras que la Figura 3.8 presenta un ejemplo equivalente para BCData.



(a) Resultados de SHIDC-B-Ki-67 HD.



(b) Resultados de SHIDC-B-Ki-67 SD.

Figura 3.7: Comparación del rendimiento de Cellpose 3 a diversos ajustes de parámetros para el conjunto de imagens SD y HD.

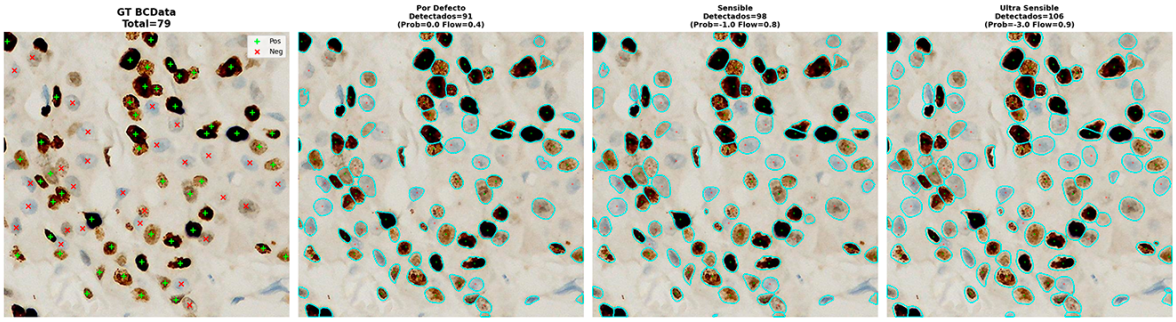


Figura 3.8: Rendimiento de cellpose 3 con BCData a diferentes parámetros.

A partir de este análisis, y considerando las características observadas en las imágenes de cada conjunto de datos, se seleccionaron configuraciones con parámetros claramente diferenciables entre sí, pero que presentaban un desempeño cualitativo comparable. La selección se realizó de modo que las configuraciones no fueran excesivamente estrictas, evitando la omisión evidente de células positivas, ni excesivamente permisivas, evitando la segmentación de regiones que correspondieran claramente a tejido o ruido.

Bajo estos criterios, se seleccionaron tres configuraciones representativas para cada conjunto de datos. El desempeño cuantitativo asociado a estas configuraciones se presenta en la Tabla 3.1.

Tabla 3.1: Evaluación de métricas de segmentación en los datasets SHIDC-B-Ki-67 y BC-Data.

Dataset	Config	Prob	Flow	D.min	D.eval	GT Rec.	GT Lost	FP	Prec.	G. Rec.
SHIDC-SD	Base	-1.0	0.8	7	15	94189	19039	72947	56.35	83.19
	Agresivo Mix	-3.0	0.8	5	15	97967	15261	74719	56.73	86.52
	Ultra Sensible	-3.0	0.9	5	15	104061	9167	104744	49.84	91.90
SHIDC-HD	Base	-1.0	0.8	30	72	106159	7069	154520	40.72	93.76
	Agresivo Mix	-3.0	0.8	30	72	108454	4774	166872	39.39	95.78
	Ultra Sensible	-3.0	0.9	35	72	110256	2972	205122	34.96	97.38
BCData	Por Defecto	0.0	0.4	20	47	71911	21927	23078	75.70	76.63
	Base	-1.0	0.8	20	47	77387	16451	31147	71.30	82.47
	Ultra Sensible	-3.0	0.9	20	47	83276	10562	38966	68.12	88.74

En esta etapa, la métrica relevante fue el *recall*. Este criterio se utilizó no solo con el objetivo de maximizar la cantidad de parches disponibles para entrenar el modelo posterior, sino también para analizar qué configuraciones de Cellpose resultan viables al momento de integrarse en el pipeline final de inferencia. En particular, se buscó asegurar que las células relevantes estuvieran presentes en la salida del segmentador, aun cuando esto implicara una disminución

en la precisión.

En escenarios donde una configuración presenta alto *recall* y baja *precisión*, se asume la presencia de un mayor número de detecciones correspondientes a fondo. Lejos de ser descartadas, estas detecciones son aprovechadas durante el entrenamiento del clasificador, permitiendo que este aprenda a identificarlas y eliminarlas en etapas posteriores del pipeline. Este enfoque resulta especialmente pertinente considerando que, para todos los conjuntos de datos evaluados excepto BCData, los valores de precisión son bajos de forma consistente.

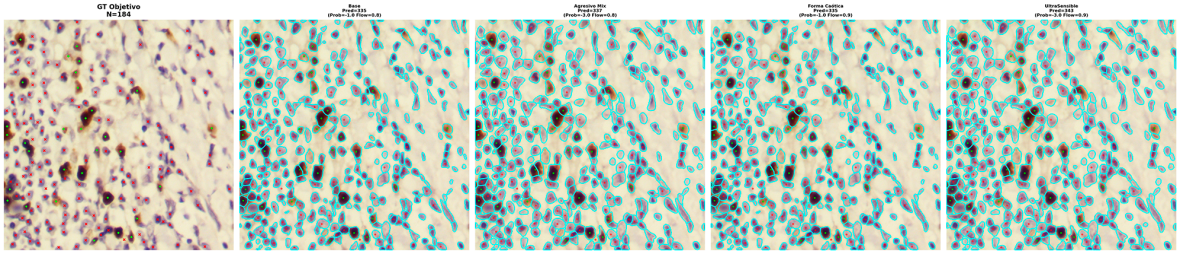
Bajo este criterio, el entrenamiento del clasificador no se basa únicamente en parches extraídos a partir de las anotaciones de verdad de terreno, sino también en parches que representan fielmente la salida real del segmentador. Esto permite que el modelo aprenda a operar sobre el mismo tipo de entradas que recibirá durante la inferencia, incorporando las características y errores propios de Cellpose.

En consecuencia, la métrica de *precisión* deja de ser prioritaria en la etapa de minería de datos. En la medida en que la clase de fondo sea correctamente aprendida, el clasificador puede filtrar eficazmente las detecciones no relevantes producidas por el segmentador, reduciendo su impacto en el análisis final. En todos los casos evaluados, la configuración denominada *Ultra Sensible* presentó el mayor valor de *recall*, por lo que fue seleccionada para esta etapa.

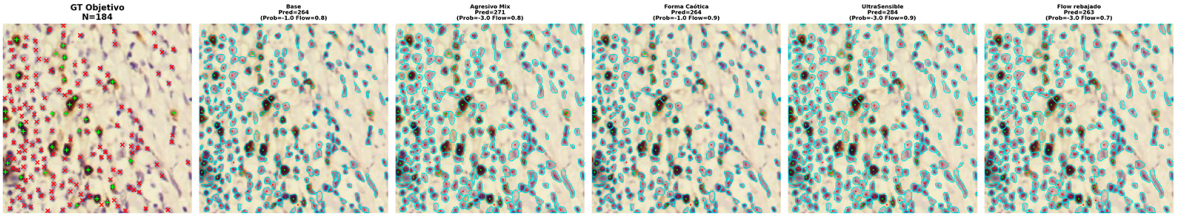
3.2.2 Cellpose SAM

Cellpose-SAM introduce un cambio de paradigma respecto a Cellpose 3, ya que dispone de un único modelo base. En consecuencia, el análisis se centró en la evaluación cualitativa del efecto de los parámetros de segmentación y en la comparación cuantitativa final.

De manera análoga al caso anterior, se analizaron distintas configuraciones de parámetros para las variantes HD y SD de SHIDC-B-Ki-67, así como para BCData. Los resultados cualitativos de estas evaluaciones se muestran en las Figuras 3.9 y 3.10. Posteriormente, se realizó una evaluación cuantitativa del desempeño de las configuraciones seleccionadas, cuyos resultados se presentan en la Tabla 3.2.



(a) Resultados de SHIDC-B-Ki-67 HD.



(b) Resultados de SHIDC-B-Ki-67 SD.

Figura 3.9: Comparación del rendimiento de Cellpose SAM a diversos ajustes de parámetros para el conjunto de imagens SD y HD.

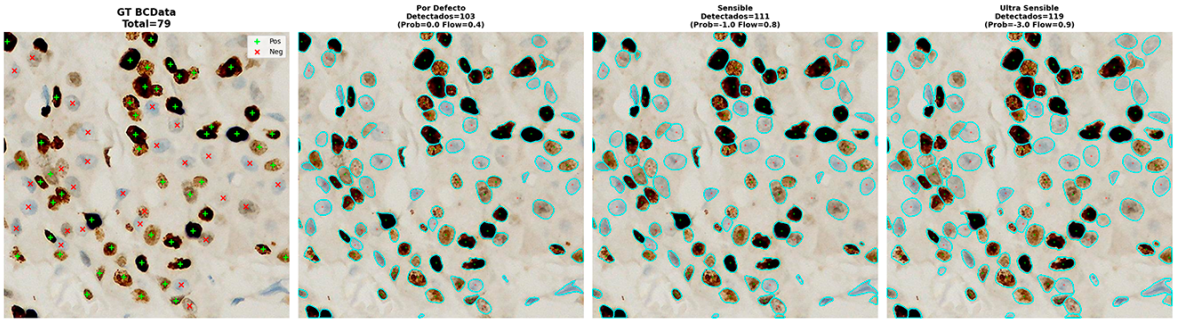


Figura 3.10: Rendimiento de cellpose SAM con BCData a diferentes parámetros.

Tabla 3.2: Evaluación de métricas de segmentación en los datasets SHIDC y BCData.

Dataset	Config	Prob	Flow	D.min	D.eval	GT Rec.	GT Lost	FP	Prec.	G. Rec.
SHIDC-SD	Base	-1.0	0.8	7	15	103770	9458	83196	55.50	91.65
	Ultra Sensible	-3.0	0.9	5	15	109461	3767	108565	50.21	96.67
SHIDC-HD	Base	-1.0	0.8	30	72	110800	2428	161426	40.70	97.86
	Ultra Sensible	-3.0	0.9	35	72	112363	865	190085	37.15	99.24
BCData	Por Defecto	0.0	0.4	20	47	84137	9701	40849	67.32	89.66
	Base	-1.0	0.8	20	47	88184	5654	50077	63.78	93.97
	Ultra Sensible	-3.0	0.9	20	47	90966	2872	57100	61.44	96.94

Puede apreciarse que Cellpose-SAM alcanza valores de *recall* superiores a los obtenidos por Cellpose 3 en los tres conjuntos de datos. En consecuencia, Cellpose-SAM fue seleccionado

como el modelo de segmentación para la etapa de minería de datos y generación de parches.

Para SHIDC-B-Ki-67 se seleccionó la configuración *Ultra Sensible*, priorizando el máximo *recall*, dado que las precisiones obtenidas eran bajas y similares entre configuraciones. En este contexto, una mayor permisividad permite asegurar la recuperación de células relevantes, aun a costa de incorporar detecciones adicionales que posteriormente pueden ser descartadas por el clasificador.

En el caso de BCData se optó por la configuración *Base*. A diferencia de SHIDC-B-Ki-67, este conjunto de datos presenta precisiones considerablemente más altas y valores de *recall* ya elevados. Bajo estas condiciones, maximizar el *recall* deja de ser un criterio prioritario, ya que configuraciones más permisivas incrementarían principalmente el número de detecciones de fondo, sin un beneficio proporcional en la recuperación de células.

Dado el comportamiento más equilibrado del segmentador en BCData, se privilegió una configuración que reflejara de forma más realista la salida esperada durante la etapa de inferencia. Esta decisión permite evitar una sobreproducción innecesaria de parches, manteniendo un compromiso adecuado entre recuperación de células, control del ruido y eficiencia computacional.

3.2.3 Minería de datos

Basándose en el trabajo de Liu et al., se adopta un enfoque de detección basado en parches clasificados como positivos y negativos. En dicho trabajo se proponen parches de tamaño 64×64 píxeles, con el objetivo de encapsular adecuadamente tanto células de menor como de mayor tamaño, minimizando el espacio libre en las primeras y evitando un recorte excesivo en las segundas.

Siguiendo este criterio, se realizaron estimaciones del diámetro celular para cada conjunto de datos. A partir de estas estimaciones, se definieron las configuraciones correspondientes tanto para el modelo de segmentación Cellpose como para el tamaño de los parches utilizados. Para el dataset de SHIDC-B-Ki-67 SD se usaron parches de 19x19 píxeles, para el dataset de SHIDC-B-Ki-67 HD parches de 72x72 píxeles y para BCData se usaron parches de 50x50. Para resoluciones bajas se optó por usar parches con una holgura de tamaño, con el fin de

poder capturar la diferencia con el fondo para la mayor cantidad de células posibles.

Para cada conjunto de datos, se recorrieron sus anotaciones y respectivas imágenes. En el caso de SHIDC-B-Ki-67, se utilizó únicamente el subconjunto de entrenamiento, mientras que para BCData se emplearon los subconjuntos de entrenamiento y validación. Las anotaciones de BCData se encontraban almacenadas en archivos `.h5`, donde cada coordenada indicaba si el núcleo correspondía a una célula positiva o negativa. En el caso de SHIDC-B-Ki-67, las anotaciones se encontraban en formato `JSON`, incluyendo una coordenada y una etiqueta con valores 1, 2 o 3, correspondientes a células positivas, negativas o linfocitos infiltrantes del tumor (TIL), respectivamente. Para efectos de este trabajo, la clase TIL fue ignorada.

El procedimiento de generación de parches siguió la siguiente estructura:

1. A partir de cada carpeta del conjunto a procesar, se cargó la imagen junto con sus anotaciones.

2. Se aplicó el modelo Cellpose para segmentar la imagen y obtener las máscaras de los núcleos detectados.

3. Para cada coordenada anotada, se verificó si existía una máscara que contuviera dicho punto.

4. En caso afirmativo, la máscara se etiquetó como positiva o negativa según correspondiera y, utilizando la coordenada como centro, se extrajo un parche. Este se almacenó en el directorio `/Train` o `/Validation`, según el caso, y en los subdirectorios `/Pos` o `/Neg`.

5. En caso de no existir una coordenada dentro de la máscara, se calculó el centroide de esta y se generó un parche utilizando dicho punto como centro. Este parche se almacenó en el directorio `/BG`, correspondiente a la clase de fondo.

6. Para cada parche generado, se registró si correspondía a un verdadero positivo (TP) o un falso positivo (FP) en relación con la segmentación realizada por Cellpose.

7. Los falsos negativos (FN) también fueron almacenados, aun cuando no hayan sido detectados por Cellpose, utilizando la coordenada anotada como centro del parche.

8. Finalmente, se recorrió la totalidad del conjunto de datos para generar todos los parches necesarios.

Este procedimiento permite evaluar de forma explícita el desempeño del modelo Cellpose en la etapa de segmentación, comparando sus detecciones con las anotaciones de referencia. Si bien el objetivo es asegurar que todos los parches correspondientes al *Ground Truth* sean incorporados al entrenamiento, resulta relevante cuantificar qué tan frecuentemente Cellpose logra segmentar correctamente los núcleos anotados.

Asimismo, este análisis permite utilizar el modelo Cellpose como generador de ejemplos de la clase de fondo cuando corresponde. Si bien una célula anotada que no es segmentada puede constituir un falso negativo, el segmentador también puede detectar estructuras que no corresponden a núcleos relevantes. Por ello, resulta necesario contar con ejemplos que representen aquellas regiones que Cellpose identifica como células, pero que, según las anotaciones de referencia, corresponden a fondo.

Bajo este enfoque, las anotaciones del *Ground Truth* o correspondientes a células positivas y negativas que caen dentro de una máscara segmentada se etiquetan como tales, mientras que aquellas máscaras que no contienen coordenadas anotadas se consideran ejemplos de la clase de fondo. De este modo, se obtienen etiquetas consistentes para las clases positiva, negativa y fondo, alineadas con el comportamiento real del segmentador dentro del pipeline propuesto.

Cabe destacar que, para el posterior análisis durante la etapa de entrenamiento, es necesario contar tanto con un conjunto de entrenamiento como con uno de validación. Dado que SHIDC-B-Ki-67 no provee explícitamente un conjunto de validación, se realizó una partición aleatoria de los parches generados, separando el conjunto original en un 80% para entrenamiento y un 20% para validación, organizados en carpetas independientes. En la tabla 3.3 se presenta el total de parches para cada clase, donde se evidencia desde un inicio un desbalance entre clases que será tratado posteriormente.

Una vez construidos los conjuntos de datos, se procede al entrenamiento del modelo de clasificación.

Tabla 3.3: Número de parches por clase para cada *Dataset*.

Dataset	Conjunto	Fondo	Positivo	negativo
SHIDC-SD	Train	89225	28084	60008
	Validation	22304	7022	15002
SHIDC-HD	Train	154529	28084	60008
	Validation	38633	7022	15002
BCData	Train	50071	33058	60780
	Validation	5223	7701	14103

3.3 Entrenamiento de modelos

Para la etapa de entrenamiento se seleccionaron arquitecturas que presentaran un desempeño sólido en tareas de detección y clasificación de núcleos celulares. En trabajos previos, como los de Liu et al., Anglada-Rotger et al. y el modelo HoVerNet, se emplean redes residuales o variantes de estas. En base a ello, el primer candidato considerado fue ResNeXt.

Otros modelos candidatos fueron seleccionados a partir del trabajo de Jeevan et al. [31], donde se comparan diversas arquitecturas del estado del arte en múltiples conjuntos de datos. En dicho estudio, ConvNeXt Tiny [44] presentó el mejor rendimiento promedio general. Para casos específicos de interés médico, como el conjunto de datos BreakHis [56], el modelo WaveMix obtuvo el mejor desempeño, seguido por EfficientNet V2-S. Sin embargo, dado que WaveMix no se encuentra disponible de forma nativa en PyTorch y requiere dependencias externas, se decidió no incluirlo en esta comparación, con el objetivo de mantener un flujo de trabajo simple, reproducible y de fácil modificación.

Finalmente, se seleccionaron tres arquitecturas base para el entrenamiento: ConvNeXt Tiny, ResNeXt-50 32x4d y EfficientNet V2-S.

El entrenamiento y la validación se realizaron a partir de los parches generados en la etapa de minería de datos. Con cada parche clasificado en una de tres clases: positivo, negativo o fondo. Para cada arquitectura las imágenes se reescalaron a 224x244 dado que los modelos emplean pesos, por defecto, pre-entrenados originalmente en Imagenet [14]. Por otro lado, no hubo variación de matiz en las imagenes para no alterar la característica Celeste-Marrón de H-DAB. Los hiperparámetros de entrenamiento se definieron en base a la documentación oficial

y a trabajos previos, manteniéndose constantes para los tres conjuntos de datos utilizados. Estas configuraciones se presentan en la Tabla 3.4.

Tabla 3.4: Resumen de hiperparámetros y configuración de entrenamiento.

Modelo	Optimizador	Learning Rate	Weight Decay	Batch Size
EfficientNet V2-S	AdamW	1×10^{-4}	1×10^{-4}	32
ResNeXt-50	AdamW	1×10^{-4}	1×10^{-4}	32
ConvNeXt Tiny	AdamW	5×10^{-5}	1×10^{-4}	32

El entrenamiento se llevó a cabo utilizando el optimizador AdamW, con un *weight decay* de 1×10^{-4} para todos los modelos. Para EfficientNet V2-S y ResNeXt-50 se empleó una tasa de aprendizaje inicial de 1×10^{-4}, mientras que para ConvNeXt Tiny se utilizó una tasa de 5×10^{-5}. El tamaño de lote fue fijado en 32 para todos los casos.

Debido al desbalance existente entre las clases, se utilizó una función de pérdida ponderada. En particular, se empleó la función de *Class-Weighted Cross-Entropy*, siguiendo el enfoque propuesto por Cortes et al. [12], cuya formulación se presenta en la Ecuación 3.5:

$$\ell_{\mathrm{WCE}}(h, x, y) = -\frac{1}{p(y)} \log\left(\frac{e^{h(x,y)}}{\sum_{y' \in \mathcal{Y}} e^{h(x,y')}}\right) \tag{3.5}$$

Este enfoque permite mitigar el sesgo hacia la clase de fondo, evitando que el modelo priorice dicha clase únicamente por su mayor frecuencia en el conjunto de datos. Los pesos obtenidos para cada entrenamiento se presentan en la Tabla 3.5.

Tabla 3.5: Pesos de clase asignados por dataset para el manejo del desequilibrio.

Dataset	Peso BG	Peso Neg	Peso Pos
BCData	0.9580329	0.7892344	1.4510759
SHIDC-B-Ki-67 SD	0.6624339	0.9849631	2.1046028
SHIDC-B-Ki-67 HD	0.5233559	1.3477148	2.8797061

Para la selección de la mejor época de entrenamiento, se consideró que, desde un punto de vista diagnóstico, resulta más relevante una correcta clasificación entre células positivas y negativas que el desempeño sobre la clase de fondo. Un buen rendimiento en *recall* y *accuracy* para las clases positiva y negativa implica una estimación más confiable del índice de proliferación.

En consecuencia, además de las métricas globales de pérdida y exactitud, se registró en cada época el valor del *F1-score* para las clases positiva y negativa. A partir de estas métricas, se implementó un mecanismo de *early stopping* junto con un programador de tasa de aprendizaje *ReduceLROnPlateau* con un factor de reducción de 0.5.

Durante el entrenamiento, si la pérdida de validación no mejoraba durante tres épocas consecutivas, la tasa de aprendizaje se reducía a la mitad. En caso de no observarse mejoras durante siete épocas consecutivas, el entrenamiento se detenía. Para considerar una mejora válida, se exigió una disminución mínima de 0.001 respecto del mejor valor registrado de pérdida de validación.

Cabe destacar que un error de clasificación entre células positivas y negativas afecta directamente el desempeño de ambas clases. Por ejemplo, una célula positiva clasificada como negativa constituye simultáneamente un falso negativo para la clase positiva y un falso positivo para la clase negativa, impactando directamente la estimación del índice de proliferación.

Para cada entrenamiento se almacenaron tanto el modelo correspondiente a la mejor pérdida de validación como aquel asociado al mejor *F1-score* promedio entre la clase positiva y negativa para la etapa de validación. Adicionalmente, se guardaron puntos de control cada cinco épocas. No obstante, para los análisis posteriores se priorizó el modelo seleccionado según el *F1-score*.

En la Tabla 3.6 se presentan los resultados obtenidos para la mejor época, en términos de *F1-score* promedio de clase positiva y negativa, de cada arquitectura en los distintos conjuntos de datos. Sin embargo, estos resultados no son suficientes para determinar un modelo ganador. Para ello, es necesario evaluar el desempeño de cada modelo dentro de un pipeline completo, incorporando la segmentación previa mediante Cellpose y utilizando el subconjunto de prueba de cada conjunto de datos.

Tabla 3.6: Desempeño para mejor época para cada modelo bajo cada conjunto de datos.

Dataset	Arch	Época	Train Loss	Train Acc. %	Val. Loss	Val. Acc. %	F1-posneg	F1-BG	F1-Macro all classes
SHIDC-SD	EfficientNet V2-S	17	0,235	88,350	0,346	85,996	0,858	0,868	0,861
	ResNeXt-50	31	0,252	87,290	0,373	84,588	0,848	0,850	0,849
	ConvNeXt Tiny	21	0,232	88,397	0,359	85,752	0,858	0,864	0,860
SHIDC-HD	EfficientNet V2-S	14	0,184	91,014	0,211	91,651	0,882	0,938	0,901
	ResNeXt-50	32	0,161	92,048	0,243	90,482	0,872	0,928	0,891
	ConvNeXt Tiny	27	0,134	93,357	0,227	91,826	0,884	0,940	0,903
BCData	EfficientNet V2-S	7	0,265	88,613	0,275	89,447	0,926	0,761	0,871
	ResNeXt-50	18	0,216	90,778	0,300	88,685	0,921	0,769	0,870
	ConvNeXt Tiny	16	0.193	91.672	0.306	89.181	0.922	0.777	0.874

3.3.1 Pruebas y elección del umbral de fondo

Puede apreciarse en la Tabla 3.6 que, para los conjuntos de datos SHIDC-B-Ki-67, el *F1-score* de la clase fondo presenta un desempeño comparable, e incluso superior, al *F1-score* promedio de las clases positiva y negativa. Este comportamiento puede atribuirse al desbalance existente entre los parches de fondo y los parches correspondientes a células positivas y negativas (véase Tabla 3.3), lo que tiende a favorecer que el modelo priorice la clasificación como fondo.

Dado que, en el pipeline propuesto, las máscaras clasificadas como fondo son descartadas en etapas posteriores, resulta crítico controlar la presencia de falsos negativos en las clases de interés, ya que estos afectan directamente la estimación final del índice de proliferación.

Por este motivo, se realizó un análisis adicional orientado a determinar cuán estricta debe ser la clasificación de una máscara como fondo para que esta sea efectivamente ignorada. El objetivo de este análisis es controlar el nivel de confianza requerido para descartar una detección, evitando así la eliminación incorrecta de núcleos positivos o negativos.

Este análisis se llevó a cabo utilizando la mejor época de entrenamiento de cada modelo y considerando los tres conjuntos de datos evaluados, evaluando cada cada modelo entrenado en cada dataset su desempeño en la contraparte de test de su dataset. Para cada imagen del subconjunto de prueba, se obtuvieron máscaras mediante Cellpose-SAM. A partir de estas máscaras se generaron parches, los cuales fueron clasificados como positivos, negativos o fondo por el modelo entrenado.

El tamaño de los parches se mantuvo constante, mientras que las configuraciones de evaluación del clasificador se ajustaron para aplicar un criterio más estricto sobre la clase de fondo, como se detalla en la Tabla 3.7. Además, se introdujo un umbral sobre la probabilidad asociada a la clase *background*. Si la probabilidad de fondo superaba dicho umbral, la máscara era descartada. En caso contrario, la máscara era clasificada como positiva o negativa, según la mayor probabilidad restante.

La diferencia entre *batchsize* y numero de *workers* se debe a limitaciones de la memoria del equipo que se estaba trabajando.

Para cada predicción positiva o negativa, se verificó si la máscara correspondiente contenía

Tabla 3.7: Configuración para inferencia en los 3 *Datasets*.

Dataset	Diámetro	Parche	Flow	Prob	Diámetro mín.	Batch Size	Num. Workers
SHIDC SD	15	19	0.8	-1.0	5	32	0
SHIDC HD	72	72	0.8	-1.0	25	16	0
BCData	47	50	0.8	-1.0	20	64	2

una coordenada anotada en el conjunto *Ground Truth*. Si la máscara no contenía ninguna coordenada, la predicción se consideró un falso positivo para la clase asignada. En caso de contener una coordenada asociada a una clase distinta, se contabilizó simultáneamente como un falso positivo para la clase predicha y un falso negativo para la clase real. Las anotaciones de *Ground Truth* no asociadas a ninguna máscara, así como las máscaras clasificadas como fondo, se contabilizaron como falsos negativos. De esta forma, tanto los fallos de cellpose como los del clasificador son penalizados.

Cabe destacar que la clasificación de fondo cumple únicamente el rol de descartar máscaras del procesamiento posterior. No participa directamente en el conteo de células positivas o negativas, sino que actúa como un mecanismo de filtrado para mejorar la robustez del sistema.

Este análisis se repitió para distintos valores del umbral de fondo, específicamente: 0.50, 0.60, 0.70, 0.75, 0.80, 0.82, 0.84, 0.85, 0.86, 0.88, 0.90, 0.92, 0.95, 0.98 y 0.99.

A partir de este barrido de umbrales, se obtuvieron los mejores valores de *F1-score* para cada modelo, junto con métricas adicionales como el error absoluto medio (MAE) del índice Ki-67 por imagen y el coeficiente de correlación de Pearson entre el índice Ki-67 estimado y el de referencia, que serán analizadas posteriormente, dichas métricas se presentan en la tabla 4.9.

La Tabla 3.8 presenta el modelo con mejor desempeño para cada conjunto de datos. En base a estos resultados, y considerando su desempeño global, se seleccionó el modelo ConvNeXt Tiny entrenado sobre BCData para su utilización en el prototipo de interfaz final.

3.4 Diseño de interfaz

La interfaz implementada replica el pipeline propuesto en este trabajo, permitiendo al usuario ejecutar de forma interactiva las etapas de segmentación y clasificación de núcleos celulares.

Tabla 3.8: Mejor modelo y umbral en los 3 *Datasets*.

Dataset	Mejor modelo	Umbral	F1-posneg
SHIDC SD	ConvNeXt Tiny	0,82	0,7557
SHIDC HD	ConvNeXt Tiny	0,92	0,7526
BCData	ConvNeXt Tiny	0.98	0.8482

El sistema permite seleccionar una carpeta correspondiente a un paciente y, mediante la configuración de parámetros como el tamaño de parche, el diámetro celular, el umbral de fondo y el modelo a utilizar, procesar una imagen individual o la totalidad de las imágenes asociadas al paciente.

Durante la ejecución, la interfaz segmenta las imágenes utilizando Cellpose-SAM y clasifica las máscaras obtenidas como células positivas, negativas o fondo. Como resultado, se entrega para cada imagen el número de células positivas y negativas, así como el índice de proliferación Ki-67 estimado.

La aplicación permite exportar los resultados del paciente en formato `CSV`, incluyendo las coordenadas y la clasificación de cada célula detectada. Asimismo, se generan máscaras binarias y visualizaciones superpuestas por imagen, facilitando su análisis posterior.

La interfaz incluye controles para visualizar u ocultar las máscaras de segmentación, así como los núcleos clasificados como positivos y negativos de forma independiente. Adicionalmente, se incorpora funcionalidad de zoom mediante la rueda del mouse, permitiendo una inspección detallada de las imágenes procesadas.

La aplicación detecta automáticamente la disponibilidad de una GPU compatible con CUDA y, en caso de estar disponible, utiliza aceleración por hardware para la inferencia. En ausencia de una GPU, el sistema opera utilizando CPU. Para el desarrollo de la interfaz se optó por configurarla en Windows. En la Figura 3.11 se muestra la interfaz gráfica por defecto del sistema, junto con una anotación numérica que identifica los distintos componentes y controles disponibles para el usuario.

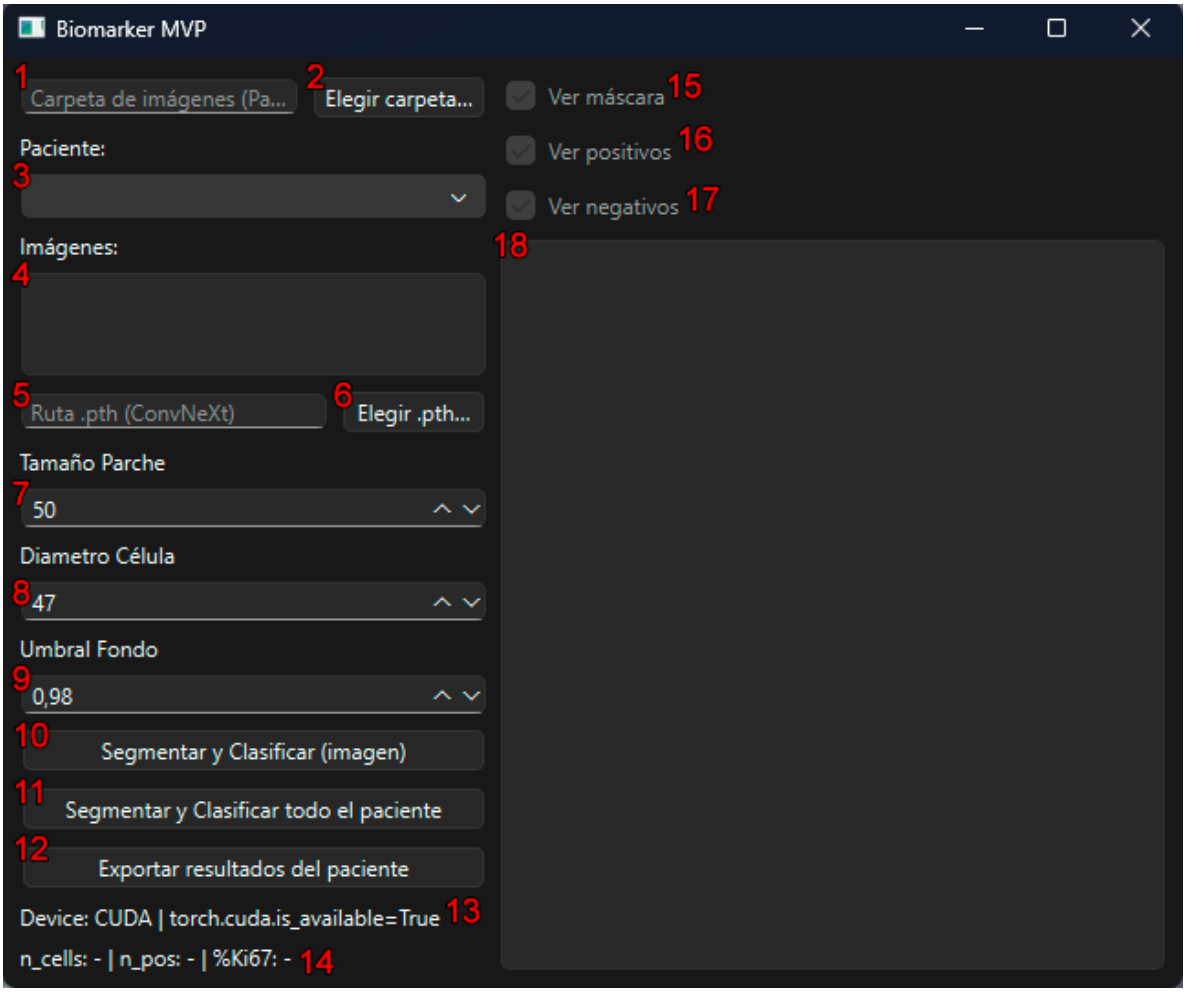


Figura 3.11: Interfaz gráfica del sistema con anotación de sus principales componentes.

La numeración presente en la figura 3.11 corresponde a los siguientes elementos:

1. Carpeta del paciente seleccionada.
2. Botón para seleccionar una carpeta desde el explorador de archivos.
3. Lista desplegable de pacientes detectados en la carpeta seleccionada.
4. Lista de imágenes asociadas al paciente seleccionado.
5. Ruta a los pesos del modelo de inferencia cargado.
6. Botón para seleccionar el archivo `.pth` del modelo desde el explorador de archivos.

7. Tamaño de parche utilizado durante la etapa de inferencia.
8. Diámetro celular utilizado para la segmentación mediante Cellpose.
9. Umbral de fondo utilizado durante la inferencia.
10. Botón para segmentar y clasificar la imagen actualmente seleccionada.
11. Botón para segmentar y clasificar todas las imágenes del paciente.
12. Botón para exportar los resultados del paciente, incluyendo máscaras y coordenadas.
13. Indicador del dispositivo de cómputo detectado (CPU o CUDA).
14. Conteo de células positivas, negativas e índice de proliferación Ki-67 estimado.
15. Casilla para visualizar las máscaras de segmentación sobre la imagen.
16. Casilla para visualizar los centroides de células positivas.
17. Casilla para visualizar los centroides de células negativas.
18. Ventana principal de visualización de la imagen seleccionada.

Una imagen de la interfaz en funcionamiento se presenta en la Figura 3.12. La imagen utilizada para esta prueba pertenece al conjunto de datos de *DeepSlides* [54].

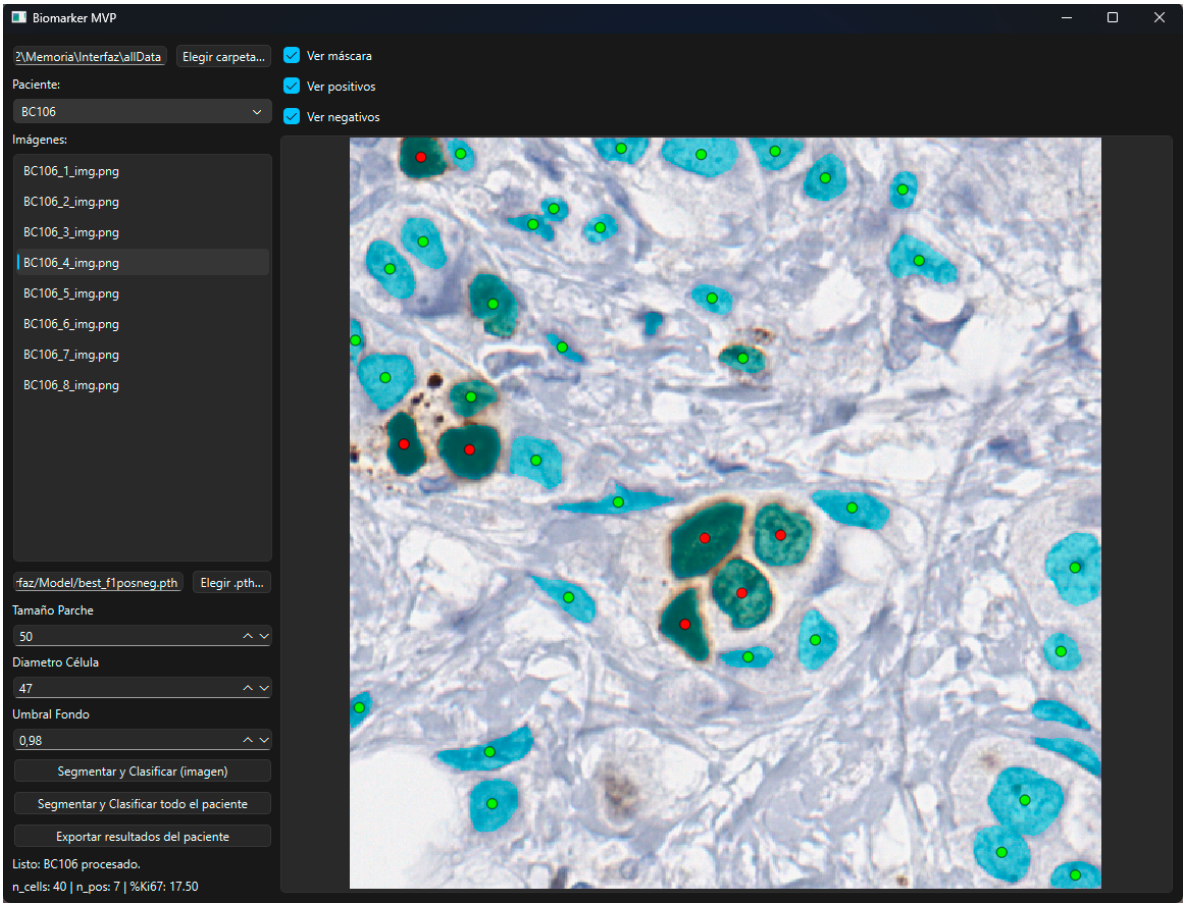


Figura 3.12: Interfaz gráfica con imagen de prueba.

La interfaz fue desarrollada utilizando PySide6 para la construcción de la interfaz gráfica y PyTorch para la ejecución del modelo de clasificación. El entorno de ejecución fue gestionado mediante Miniconda, utilizando un entorno dedicado denominado `biomarker_mvp`. La ejecución de este programa en *Windows* viene asociado con un .bat.

El código de la interfaz, las dependencias, instrucciones y requerimientos, junto con los *scripts* utilizados para el entrenamiento y evaluación de los modelos, se encuentra disponible en el repositorio asociado a este trabajo 5.2.

4 Resultados y Análisis

En esta sección se presenta un desglose de los resultados obtenidos, específicamente del rendimiento final de los modelos ya entrenados. A su vez, se compara su desempeño con otros trabajos de la literatura y se discuten las razones de las diferencias.

En esta etapa, adicionalmente, se emplean métricas orientadas a evaluar la calidad de la cuantificación del índice Ki-67 a nivel de imagen. A diferencia de las métricas de clasificación nuclear definidas previamente, estas métricas permiten analizar el error global en la estimación del biomarcador y la consistencia de la tendencia entre valores predichos y de referencia. En particular, se utilizan el error absoluto medio (MAE), el error cuadrático medio (MSE), la raíz del error cuadrático medio (RMSE) y el coeficiente de correlación de Pearson.

El error absoluto medio (*Mean Absolute Error*, MAE) mide la magnitud promedio del error entre el índice Ki-67 estimado y el valor de referencia, sin considerar la dirección del error:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{4.6}$$

El error cuadrático medio (*Mean Squared Error*, MSE) penaliza de forma más severa los errores grandes, al elevar al cuadrado la diferencia entre la predicción y el valor real:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{4.7}$$

La raíz del error cuadrático medio (*Root Mean Squared Error*, RMSE) corresponde a la raíz cuadrada del MSE y se expresa en las mismas unidades que el índice Ki-67, facilitando su interpretación:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{4.8}$$

Finalmente, el coeficiente de correlación de Pearson evalúa la relación lineal entre los valores estimados y los valores de referencia del índice Ki-67, indicando si el modelo preserva la

tendencia relativa entre imágenes con distintos niveles de proliferación:

$$Pearson(\rho) = \frac{\sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{4.9}$$

Donde y es el dato real y $\hat{y}$ es la predicción.

Estas métricas se utilizan de manera complementaria para caracterizar tanto la magnitud del error en la estimación del índice Ki-67 como la coherencia global de las predicciones del modelo respecto a los valores de referencia.

4.1 Desempeño de los modelos ganadores

En la Tabla 4.9 se presenta la comparación de los umbrales óptimos seleccionados para cada modelo y conjunto de datos, definidos a partir del *F1-score* promedio entre las clases Ki-67 positiva y Ki-67 negativa. Esta métrica se utilizó como criterio principal de selección, ya que permite evaluar de forma equilibrada la capacidad del modelo para discriminar ambas clases nucleares, directamente relacionadas con la estimación del índice de proliferación celular. Para cada columna, el mayor valor obtenido se resalta en negrita.

Tabla 4.9: Desempeño de clasificación nuclear para los umbrales óptimos de cada modelo y dataset.

Dataset	Arquitectura	BG thr	Prec. Pos	Rec. Pos	Prec. Neg	Rec. Neg	F1 Pos	F1 Neg	F1 Pos/Neg	F1 Pos/Neg Micro
BCData	ConvNeXt-Tiny	0.98	**0.853**	0.885	0.840	**0.816**	**0.869**	**0.828**	**0.848**	**0,841**
	ResNeXt-50	0.95	0.849	0.889	0.855	0.797	0.868	0.825	0.847	0,839
	EfficientNet-V2-S	0.82	0.836	**0.897**	**0.868**	0.790	0.865	0.827	0.846	0,840
SHIDC-SD	ConvNeXt-Tiny	0.82	**0.754**	0.875	0.657	**0.752**	**0.810**	**0.701**	**0.756**	**0,736**
	ResNeXt-50	0.70	0.742	0.871	**0.665**	0.725	0.802	0.694	0.748	0,729
	EfficientNet-V2-S	0.80	0.739	**0.880**	0.652	0.745	0.803	0.695	0.749	0,730
SHIDC-HD	ConvNeXt-Tiny	0.92	0.752	0.866	0.628	**0.792**	**0.805**	0.701	**0.753**	**0,733**
	ResNeXt-50	0.82	**0.757**	0.853	0.634	0.771	0.802	0.696	0.749	0,729
	EfficientNet-V2-S	0.80	0.711	**0.890**	**0.657**	0.753	0.790	**0.702**	0.746	0,731

Adicionalmente, para cada configuración evaluada se calcularon métricas a nivel de imagen orientadas a evaluar la calidad de la cuantificación del índice Ki-67. En particular, el índice Ki-67 estimado a partir de la predicción del modelo se comparó con el valor de referencia (*ground truth*) correspondiente a cada imagen. A partir de esta comparación se obtuvieron el error absoluto medio (MAE), el error cuadrático medio (MSE) y el coeficiente de correlación de Pearson. Estas métricas se presentan en la Tabla 4.10, donde los mejores valores se indican

en negrita.

Tabla 4.10: Evaluación de la cuantificación del índice Ki-67 a nivel de imagen para los umbrales óptimos de cada modelo.

Dataset	Arquitectura	BG thr	F1 Pos/Neg	MAE	MSE	Pearson
BCData	ConvNeXt-Tiny	0.98	**0.848**	**0.051**	**0.0065**	0.960
	ResNeXt-50	0.95	0.847	0.057	0.0079	0.960
	EfficientNet-V2-S	0.82	0.846	0.062	0.0085	**0.961**
SHIDC-SD	ConvNeXt-Tiny	0.82	**0.756**	0.078	0.0173	0.867
	ResNeXt-50	0.70	0.748	0.078	**0.0162**	**0.877**
	EfficientNet-V2-S	0.80	0.749	**0.076**	0.0163	0.875
SHIDC-HD	ConvNeXt-Tiny	0.92	**0.753**	**0.065**	**0.0101**	**0.923**
	ResNeXt-50	0.82	0.749	0.066	0.0104	0.918
	EfficientNet-V2-S	0.80	0.746	0.065	0.0106	0.916

A diferencia del cálculo del *F1-score*, la *precision* y el *recall*, que se realizó de forma global agregando todas las predicciones del conjunto de evaluación, las métricas MAE, MSE y Pearson se calcularon a nivel de imagen. Esta diferencia metodológica responde a la naturaleza distinta de las preguntas que aborda cada grupo de métricas.

El *F1 Pos/Neg* corresponde al promedio aritmético entre el *F1 Pos* y el *F1 Neg* (macro average), ya que se encuentra directamente alineado con el objetivo de estimar el índice de proliferación, otorgando el mismo peso a las clases positivas y negativas. En contraste, el *F1 Pos/Neg micro* se calcula considerando conjuntamente todas las instancias positivas y negativas del conjunto de datos, siendo sensible al desbalance de clases. Esta métrica se emplea posteriormente con fines comparativos frente a otros trabajos.

Las métricas de *precision*, *recall* y *F1-score* evalúan el desempeño a nivel de instancia (núcleo), midiendo la capacidad del modelo para clasificar correctamente cada célula individual. Por esta razón, dichas métricas se calcularon agregando los verdaderos positivos, falsos positivos y falsos negativos de todas las imágenes del conjunto, de modo que cada núcleo contribuya de manera equivalente al análisis. Este enfoque evita sesgos asociados a la variabilidad en el número de anotaciones por imagen.

Si estas métricas se calcularan de forma independiente por imagen y luego se promediaran, una imagen con pocas anotaciones tendría el mismo peso que otra con un número significati-

vamente mayor de núcleos. Por ejemplo, no detectar ninguna célula en una imagen con dos anotaciones produciría un *recall* nulo, mientras que fallar dos detecciones en una imagen con veinte anotaciones resultaría en un *recall* de 0.9. Al promediar ambos valores se obtendría un *recall* de 0.45, penalizando el resultado de forma desproporcionada, aun cuando el *recall* global considerando todas las instancias sería cercano a 0.82.

Este enfoque es coherente con el análisis de desempaño del *pipeline*, resulta adecuado evaluar la calidad de la clasificación nuclear considerando todas las predicciones de forma conjunta, en lugar de fragmentar el análisis por imagen.

Por otro lado, las métricas MAE y MSE permiten evaluar la precisión de la estimación del índice Ki-67 a nivel de imagen o región de interés, independientemente de la identificación exacta de cada núcleo individual. En este contexto, un modelo puede presentar errores a nivel de instancia y, aun así, mantener una proporción adecuada entre núcleos positivos y negativos, produciendo un índice Ki-67 con bajo error respecto al valor de referencia. Este análisis permite evaluar la robustez del sistema frente a errores locales de detección o clasificación.

La correlación de Pearson complementa este análisis al evaluar la consistencia de la tendencia global del modelo, indicando si variaciones en el índice Ki-67 real se reflejan en variaciones concordantes en el índice estimado. Una correlación elevada indica que el modelo preserva la relación relativa entre imágenes con distintos niveles de proliferación, independientemente del error absoluto.

En conjunto, mientras el análisis basado en *precision*, *recall* y *F1-score* evalúa la confiabilidad de la clasificación nuclear a nivel de instancia, las métricas MAE, MSE y Pearson permiten analizar la capacidad del sistema para estimar el índice Ki-67 a nivel de imagen. Este enfoque resulta particularmente relevante considerando que el análisis parche a parche constituye una práctica habitual en patología digital, donde comúnmente se seleccionan regiones *hotspot* para el cálculo del índice de proliferación [40].

Cabe recordar que el *pipeline* no contempla el procesamiento directo de WSIs completas, ya que estas requieren una etapa previa de preprocesamiento y extracción de parches.

A partir del análisis conjunto de las Tablas 4.9 y 4.10, se observa que los modelos basados en

ConvNeXt-Tiny presentan el mejor desempeño global tanto en la clasificación nuclear como en la estimación del índice Ki-67. Las diferencias respecto a las demás arquitecturas evaluadas no superan los 0.01. En términos de MAE y MSE, ConvNeXt-Tiny obtiene los mejores resultados en dos de los tres conjuntos de datos. En cuanto a la correlación de Pearson, su mejor desempeño se observa en BCData, mientras que en SHIDC-HD el valor más alto corresponde a EfficientNet-V2-S, con una diferencia del orden de 10^{-3}.

En el conjunto BCData, ConvNeXt-Tiny presenta valores de *F1-score* superiores a los observados en los demás conjuntos de datos, junto con errores MAE y MSE menores, lo que se traduce en una estimación más precisa del índice Ki-67 a nivel de imagen. En conjunto, estos resultados respaldan la elección de ConvNeXt-Tiny como la arquitectura seleccionada para su integración en la interfaz final del sistema, al presentar el mejor equilibrio entre desempeño de clasificación nuclear y capacidad de cuantificación del biomarcador.

4.2 Comparación con otros trabajos

En esta sección se analiza el desempeño del modelo seleccionado, considerando su integración en el *pipeline* completo, y se compara con trabajos previos reportados en la literatura. El objetivo es evaluar en qué medida la propuesta se encuentra alineada con el estado del arte y justificar las decisiones metodológicas adoptadas.

Para realizar una comparación adecuada, se consideran únicamente trabajos que emplean los mismos conjuntos de datos. En particular, se incluyen PathoNet [48] y KpiNet [42], dado que ambos reportan resultados sobre los mismos *datasets* utilizados en este trabajo.

En el caso de PathoNet, los resultados disponibles corresponden al *dataset* SHIDC-B-Ki-67 utilizando parches de 256×256. Por su parte, KpiNet reporta resultados tanto para BCData como para SHIDC-B-Ki-67, también en su variante de 256×256. En consecuencia, la comparación se limita a estos conjuntos de datos y resoluciones, con el fin de asegurar condiciones equivalentes de evaluación.

4.2.1 Comparación con SHIDC-B-Ki-67 256x256

Tanto KPi-Net como PathoNet reportan métricas globales de *precision* y *recall* agregadas sobre todo el conjunto de datos, así como el error RMSE asociado a la estimación del índice Ki-67 a nivel de imagen o parche. En consecuencia, la comparación debe realizarse utilizando las métricas y configuraciones originalmente reportadas por cada trabajo.

No obstante, los resultados publicados para PathoNet presentan diferencias respecto a los reportados posteriormente por KPi-Net. En particular, PathoNet incluye en su análisis la etiqueta correspondiente a linfocitos infiltrantes tumorales (TIL), mientras que en este trabajo dicha etiqueta es ignorada. Además, el trabajo de KPi-Net introduce modificaciones metodológicas y métricas adicionales que no estaban presentes en PathoNet, sugiriendo una mejora sobre el enfoque original.

Por estas razones, y con el fin de mantener una comparación coherente y metodológicamente consistente, se opta por utilizar la versión del análisis presentada en el trabajo de KPi-Net, la cual además corresponde a un estudio más reciente. Considerando estas condiciones, y limitándose únicamente a las métricas comunes y respaldables entre los trabajos, se construye la Tabla 4.11, donde se compara el modelo ganador ConvNeXt-Tiny integrado en el pipeline propuesto con los resultados reportados en la literatura, cabe destacar que los valores en las columnas de **Promedio** corresponden a los *micro average*, intentando mantener el formato escogido por Liu et al.

Por último, se debe considerar que esta comparativa es únicamente referencial respecto a lo que se esperaría del estado del arte en la estimación del índice de proliferación Ki-67 mediante conteo de instancias. Esto se debe a que los detalles de entrenamiento, así como el código del trabajo de Liu et al., no se encuentran disponibles públicamente, por lo que una evaluación justa solo podría realizarse bajo condiciones de entrenamiento y configuraciones comunes.

4.2.2 Comparación con BCData

De manera análoga al caso de SHIDC-B-Ki-67, se construye una tabla comparativa utilizando los resultados reportados en el trabajo de Qi Liu et al.[42] y los obtenidos en este estudio. La comparación se presenta en la Tabla 4.12 y se limita a las métricas comunes entre los trabajos.

Tabla 4.11: Comparación del desempeño de cuantificación Ki-67 entre trabajos del estado del arte y el pipeline propuesto para el dataset de SHIDC-B-Ki-67 256x256. Valores para PathoNet y KPi-Net tomados de Qi Liu et al.[42] y reorganizados para fines comparativos.

	Ki67+			**Ki67-**			**Promedio**					
Modelo	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	**F1 posneg macro**	**RMSE**	**R**
PathoNet	84.76	86.58	85.73	74.94	82.01	78.75	78.56	84.05	82.26	82.24	0.04803	0.964
KPi-Net	89.45	90.95	90.19	83.67	85.23	84.99	84.35	87.09	85.79	87.59	0.04632	0.965
Pipeline propuesto	75.41	87.49	81.00	65.69	75.24	70.14	68.88	79.23	73.69	75.57	0.13159	0.867

Para este conjunto de datos, se utiliza nuevamente el modelo ConvNeXt-Tiny, seleccionado previamente como el mejor desempeño para BCData, considerando su integración completa dentro del pipeline propuesto. De este modo, la comparación refleja no solo el rendimiento del clasificador, sino también el efecto conjunto de la segmentación, la clasificación nuclear y el cálculo del índice Ki-67 bajo condiciones equivalentes de evaluación.

Tabla 4.12: Comparación del desempeño de cuantificación Ki-67 entre trabajos del estado del arte y el pipeline propuesto para el dataset de BCData. Valores para PathoNet y KPi-Net tomados de Qi Liu et al.[42] y reorganizados para fines comparativos.

	Ki67+			**Ki67-**			**Promedio**					
Modelo	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	**F1 posneg macro**	**RMSE**	**R**
PathoNet	86.77	88.32	87.68	77.87	82.33	80.15	80.32	83.43	82.92	82.91	0.05034	0.958
KPi-Net	88.78	90.57	89.56	80.86	84.35	82.94	82.82	86.46	84.25	86.25	0.04854	0.962
Pipeline propuesto	85.32	88.46	86.86	83.97	81.60	82.77	84.44	83.89	84.17	84.82	0.08076	0.960

4.2.3 Análisis y justificaciones

En el caso del dataset SHIDC-B-Ki-67, el pipeline propuesto presenta un desempeño inferior al de los modelos del estado del arte, con una diferencia cercana a los diez puntos porcentuales en las métricas reportadas. Esta disminución no puede atribuirse únicamente a la etapa de clasificación, sino también a la detección inicial de núcleos realizada por Cellpose. Si bien sería posible aumentar la permisividad del segmentador para recuperar más instancias, se optó por mantener una configuración más restrictiva con el objetivo de evitar una sobreproducción de parches. Un análisis posterior permitiría determinar si esta disminución de desempeño se encuentra asociada principalmente a la variabilidad intrínseca del dataset o a la calidad de sus anotaciones. En cualquier caso, aunque el modelo podría utilizarse de forma exploratoria en un prototipo, sus resultados no son directamente comparables con los del estado del arte para este conjunto de datos.

Para el dataset BCData, en cambio, los resultados obtenidos son más consistentes con los reportados en la literatura. El pipeline propuesto supera a PathoNet en 8 métricas de 12 y presenta un desempeño inferior a KPi-Net, aunque sin diferencias superiores a cinco puntos porcentuales en ninguno de los casos. Esto sitúa al modelo dentro de un rango comparable al estado del arte para este conjunto de datos.

Se observa, no obstante, que el RMSE del pipeline propuesto es mayor en comparación con los otros trabajos. Esto indica una mayor sensibilidad a errores de predicción puntuales, particularmente en imágenes específicas. Sin embargo, el coeficiente de correlación de Pearson se mantiene en un rango similar al de los modelos comparados, situándose entre ambos. Este resultado respalda el uso del modelo como un prototipo válido para un MVP, como es el caso de la interfaz desarrollada, con la posibilidad de reemplazar el modelo en caso de que su desempeño no resulte satisfactorio en escenarios reales.

Un RMSE relativamente alto puede resultar crítico en casos cercanos a los umbrales de decisión clínica del índice Ki-67, donde pequeñas variaciones pueden modificar la categorización del nivel de proliferación. Bajo este contexto, el sistema no busca reemplazar al patólogo, sino asistirlo, permitiendo visualizar las predicciones y reconocer posibles fuentes de error. Para el modelo entrenado en BCData, un MAE de 0.051 (véase la Tabla 4.10) indica una desviación promedio cercana al $\pm 5\%$, mientras que el RMSE sugiere que, en los peores casos, el error podría alcanzar aproximadamente un $\pm 8\%$. Si bien estos márgenes pueden resultar relevantes en escenarios limítrofes, son comparables a los reportados por otros trabajos del estado del arte, donde los errores máximos también se sitúan en torno al $\pm 5\%$. Considerando que se trata de un prototipo, este nivel de desempeño resulta aceptable para los fines planteados.

Existen métricas adicionales reportadas en la literatura que no fueron incorporadas en este análisis, como el *cutoff accuracy* de PathoNet o el *PI accuracy rate* propuesto por KPi-Net. La exclusión de estas métricas responde a limitaciones metodológicas. En el caso de PathoNet, el *cutoff accuracy* se evalúa a nivel de paciente, determinando únicamente si la predicción final se encuentra en el mismo rango clínico que el valor de referencia, sin cuantificar la magnitud del error. Este enfoque requiere anotaciones a nivel de paciente, las cuales no están disponibles en BCData, impidiendo una comparación justa. En cuanto al *PI accuracy rate* de KPi-Net, su definición no está explicitamente documentada ni tampoco su diferencia con el *cutoff accuracy*

de PathoNet, y no se especifica claramente cómo se maneja la agregación de imágenes o pacientes en BCData. Dado que no existe código público que permita reproducir o verificar esta métrica, se optó por no incluirla en el análisis.

En síntesis, el modelo entrenado sobre BCData y su integración en el pipeline propuesto no superan el estado del arte actual, pero presentan un desempeño cercano y consistente. Esto justifica su utilización dentro del prototipo desarrollado, el cual ofrece una interfaz configurable y extensible, abierta a futuras actualizaciones tanto en la etapa de segmentación como en la de clasificación, y no necesariamente limitada al biomarcador Ki-67.

4.3 Diferencias entre Datasets

Es posible apreciar en los resultados presentados en las Tablas 4.11 y 4.12 que el desempeño general sobre el dataset BCData es superior al observado en SHIDC-B-Ki-67. Una posible explicación de esta diferencia se relaciona con el estado de las anotaciones, la calidad de las imágenes y la consistencia entre muestras.

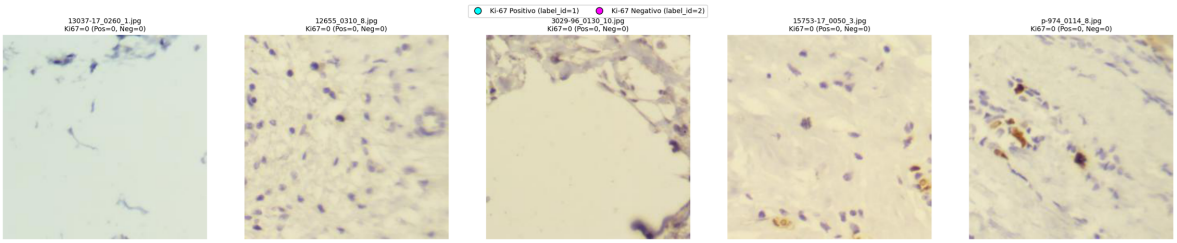
El trabajo de Liu et al.[42] señala explícitamente que el dataset SHIDC-B-Ki-67 presenta una variabilidad notable en la tinción, así como una alta superposición de células. En contraste, para BCData se reporta una mayor homogeneidad visual general, aunque con presencia de heterogeneidad en la tinción y agrupamientos celulares más severos. Estas características pueden afectar de manera distinta tanto la segmentación como la clasificación nuclear.

Por otro lado, Negahbani et al.[48] mencionan que, si bien las anotaciones de su dataset fueron realizadas por expertos, la gran cantidad de núcleos e imágenes implica que estas no están exentas de posibles errores. Este factor introduce una fuente adicional de incertidumbre en el entrenamiento y evaluación de los modelos.

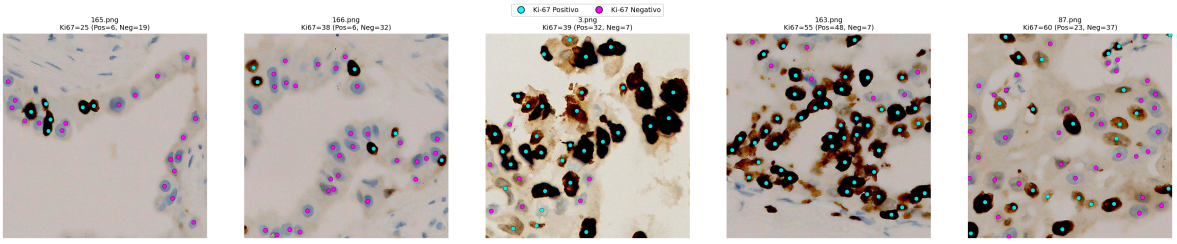
El objetivo de esta comparación entre datasets no es determinar la superioridad de un conjunto de datos sobre otro, sino evidenciar posibles factores que influyen en el desempeño observado durante las etapas de entrenamiento y evaluación. Los resultados presentados deben interpretarse de manera descriptiva, ya que emitir juicios más concluyentes requeriría la validación por parte de un patólogo experto, instancia que no estuvo disponible durante el desarrollo de este trabajo.

Con el fin de complementar el análisis cuantitativo, se realizaron dos comparaciones de carácter cualitativo. La primera consistió en identificar, dentro de cada conjunto de datos, las imágenes con menor cantidad de anotaciones de Ki-67 positivo y Ki-67 negativo.

En la Figura 4.13, la subfigura 4.13a muestra imágenes del conjunto de prueba de SHIDC-B-Ki-67 con la menor cantidad de anotaciones. En este caso, se presentan cinco de las veinte imágenes que empatan con cero anotaciones de Ki-67, tanto positivas como negativas. En contraste, la subfigura 4.13b presenta el mismo análisis para BCData, donde ninguna imagen carece de anotaciones y el mínimo observado corresponde a 25 núcleos anotados.



(a) Menor cantidad de anotaciones para SHIDC-B-Ki-67.



(b) Menor cantidad de anotaciones para BCData.

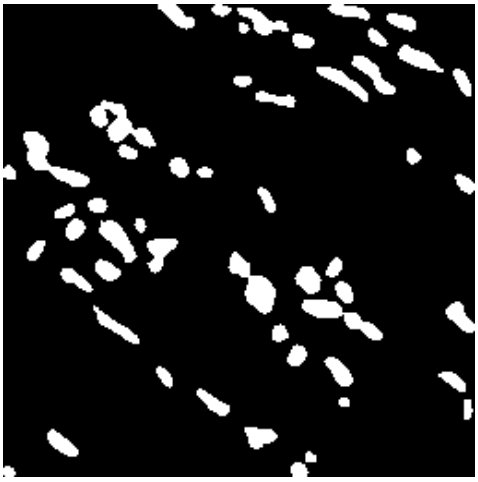
Figura 4.13: Imágenes con la menor cantidad de anotaciones para cada subconjunto de prueba de cada dataset.

Es posible apreciar que las imágenes presentadas en dichas figuras contienen una menor cantidad de tinción; sin embargo, aún se observan regiones teñidas que no se encuentran anotadas. Determinar si estas corresponden efectivamente a núcleos celulares o a otros elementos del tejido escapa al alcance de este trabajo. No obstante, un segmentador como Cellpose puede identificar y segmentar regiones no anotadas, las cuales posteriormente son enviadas a la etapa de clasificación.

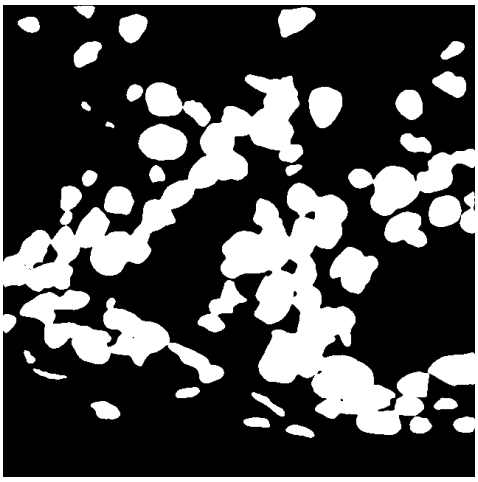
Este comportamiento se ilustra en la Figura 4.14. En particular, para la versión de Cellpose utilizada en la interfaz, la imagen `163.png` de BCData, mostrada en la Figura 4.14b, presenta

múltiples regiones segmentadas como posibles células, incluyendo estructuras y tinciones adicionales que no se encuentran anotadas. Un comportamiento similar se observa en la imagen `p-974_0114_8` de SHIDC-B-Ki-67, presentada en la Figura 4.14a, donde también se detectan máscaras ausentes en las anotaciones originales.

Cabe destacar que ninguna región es clasificada sin haber sido previamente detectada por Cellpose. La validez biológica de estas detecciones, sin embargo, depende de la evaluación de un patólogo. La misma característica que permite a Cellpose generalizar y emplearse en distintos conjuntos de datos también limita su especificidad al enfrentarse a un dataset particular, lo que puede traducirse en la detección de estructuras que no corresponden a núcleos relevantes.



(a) Máscara de segmentación de imagen p-974_0114_8.jpg de SHIDC-B-Ki-67.

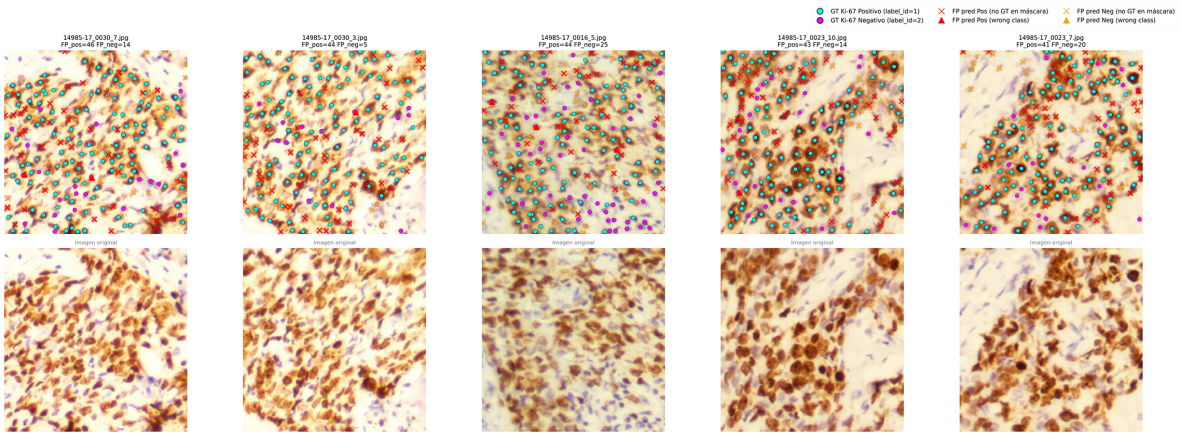


(b) Máscara de segmentación de imagen 163.png de BCData.

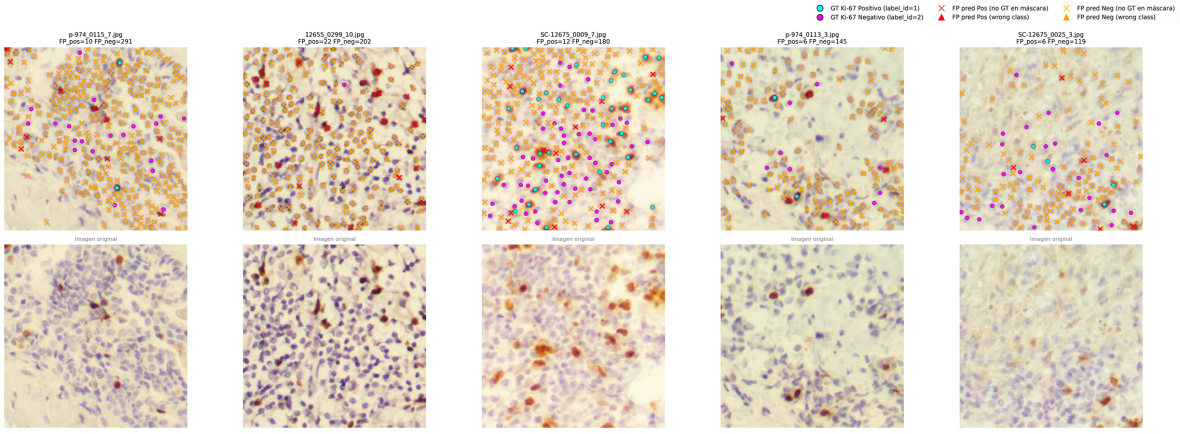
Figura 4.14: Máscaras de segmentacionde cellpose para las imágenes de prueba.

Este razonamiento se ve reforzado al analizar la distribución de falsos positivos y falsos negativos para cada conjunto de datos. En la Figura 4.15 se presentan ejemplos representativos para SHIDC-B-Ki-67. En particular, la subfigura 4.15a muestra las imágenes con mayor cantidad de falsos positivos correspondientes a Ki-67 positivo. En estas se observan detecciones en regiones que, según la verdad de terreno, no deberían ser clasificadas como positivas, pero que presentan algún tipo de tinción o estructuras de forma aproximadamente circular.

De manera análoga, la subfigura 4.15b presenta las imágenes con mayor número de falsos positivos para la clase Ki-67 negativo. En este caso, se aprecia un patrón similar, donde el segmentador identifica regiones no anotadas que poseen características visuales compatibles con núcleos, pero que no se encuentran etiquetadas en el conjunto de datos.



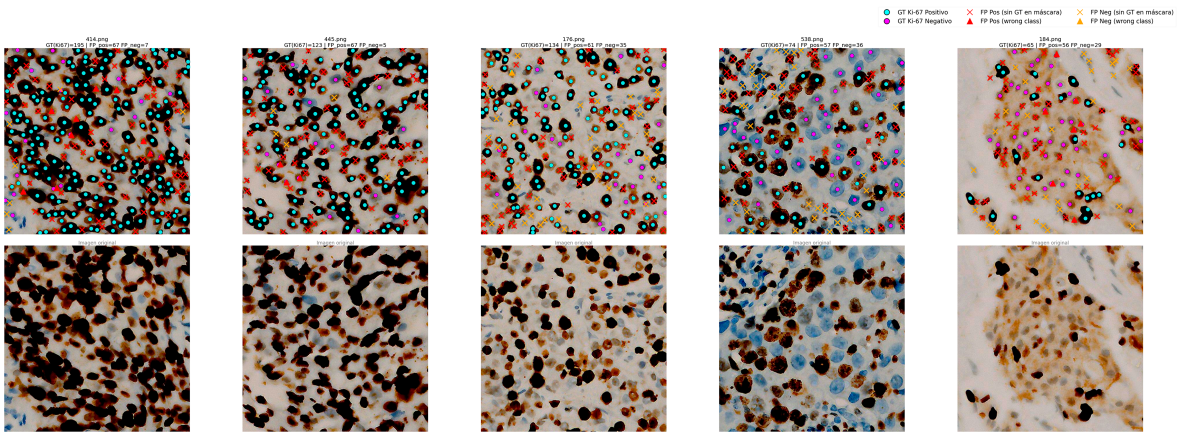
(a) 5 imágenes de SHIDC-B-Ki-67 con mayor numero de falsos positivos.



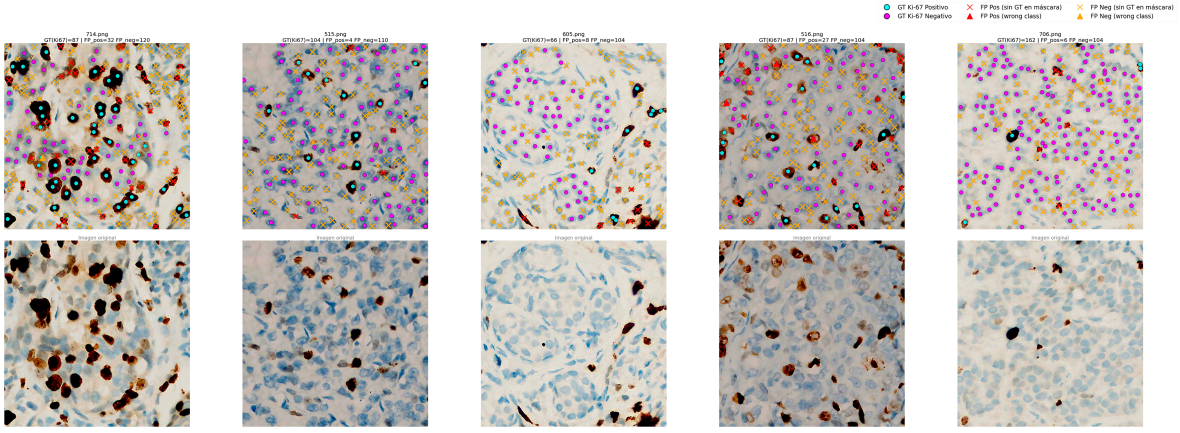
(b) 5 imágenes de SHIDC-B-Ki-67 con mayor numero de falsos negativos.

Figura 4.15: Imágenes con más falsos positivos para SHIDC.

En el caso de BCData también se observan detecciones en regiones no anotadas. Esta característica se aprecia en la Figura 4.16. En la subfigura 4.16a se muestran las imágenes con mayor número de falsos positivos para la clase Ki-67 positivo, mientras que en la subfigura 4.16b se presentan las imágenes con más falsos positivos correspondientes a Ki-67 negativo. En ambos casos, las detecciones se concentran en regiones que no cuentan con anotación en la verdad de terreno, pero que presentan patrones visuales compatibles con núcleos celulares.



(a) 5 imágenes de BCData con mayor numero de falsos positivos.



(b) 5 imágenes de BCData con mayor numero de falsos negativos..

Figura 4.16: Imágenes con más falsos positivos para BCData.

En síntesis, se observa que el pipeline tiende a detectar y clasificar elementos que presentan morfología compatible con núcleos celulares y algún grado de tinción. Determinar si estas detecciones corresponden efectivamente a células Ki-67 positivas o negativas requiere la validación de un experto. En los casos extremos, se aprecia además que las imágenes de BCData

presentan, en general, un mayor número de anotaciones de Ki-67 positivo y negativo en comparación con SHIDC-B-Ki-67.

Por otro lado, al analizar las imágenes con mayor cantidad de falsos positivos, se observa que la naturaleza de estas detecciones es similar entre ambos conjuntos de datos, considerando además que la etiqueta correspondiente a TIL en SHIDC-B-Ki-67 no fue incluida en este trabajo. Este análisis se realizó exclusivamente sobre el conjunto de prueba. Un estudio adicional sobre el conjunto de entrenamiento podría aportar mayor claridad, siempre que se realice con la supervisión de un patólogo.

A partir de este análisis cualitativo, es razonable plantear que este fenómeno, y su posible comportamiento sistemático a lo largo del conjunto de entrenamiento, podría estar influyendo en las diferencias de desempeño observadas entre los modelos entrenados con SHIDC-B-Ki-67 y BCData.

4.4 Limitaciones del prototipo

La interfaz desarrollada corresponde a un prototipo funcional orientado a la demostración del pipeline propuesto. No fue diseñada como una herramienta clínica ni validada para su uso en entornos diagnósticos reales.

En primer lugar, el sistema opera sobre imágenes individuales o conjuntos de imágenes previamente extraídas, y no sobre láminas histopatológicas completas (WSI). Por lo tanto, no aborda desafíos asociados al manejo de imágenes de gran tamaño, como la gestión de memoria o la selección automática de regiones de interés.

En segundo lugar, el prototipo asume que las imágenes de entrada corresponden a preparaciones inmunohistoquímicas comparables a aquellas utilizadas durante el entrenamiento, específicamente DAB-H. No se consideran variaciones extremas de tinción, escaneo o adquisición propias de distintos laboratorios, lo que puede afectar el desempeño del sistema. Para un uso más generalizado, serían necesarios modelos entrenados con conjuntos de datos más amplios y diversos.

Asimismo, el sistema no incorpora mecanismos de calibración clínica ni validación interobserva-

dor. El índice Ki-67 estimado se utiliza únicamente como una medida cuantitativa automática para fines de análisis y comparación experimental, y no como un reemplazo de la evaluación realizada por un patólogo.

Desde el punto de vista del modelo, la clasificación se realiza a nivel de parche y depende directamente del desempeño del segmentador. Errores en la etapa de segmentación, tales como sobresegmentación o subsegmentación, pueden propagarse al resultado final. Si bien el uso de un umbral de fondo permite mitigar parcialmente este efecto, no lo elimina por completo. De la misma forma, errores del segmentador y del clasificador se acumulan a lo largo del pipeline, generando dos posibles fuentes de error. Esta composición de errores puede hacer que el sistema sea más susceptible a desviaciones en la estimación final del índice Ki-67 en comparación con enfoques monolíticos o altamente especializados.

Adicionalmente, el segmentador Cellpose se encuentra integrado de forma fija dentro del código del prototipo. Si bien su comportamiento puede ajustarse mediante la modificación de parámetros o rutas, el sistema no permite intercambiar dinámicamente el segmentador por otro modelo sin realizar cambios directos en el código. Esta decisión responde a criterios de simplicidad y estabilidad propios de un prototipo, pero limita la flexibilidad del sistema para escenarios de experimentación más avanzada.

De manera similar, algunos elementos de la interfaz se encuentran definidos de forma específica para el biomarcador Ki-67. Si bien estos componentes son relativamente sencillos de modificar a nivel de código, al haberse trabajado exclusivamente con este biomarcador, ciertas etiquetas y configuraciones permanecen fijas.

Asimismo, la interfaz asume que los pesos utilizados corresponden a un clasificador ConvNeXt-Tiny. Si bien es posible intercambiar los pesos del modelo, la arquitectura del clasificador permanece constante. Esta decisión se adoptó con el objetivo de simplificar el diseño del prototipo y facilitar su uso como MVP, a costa de una menor generalidad arquitectónica.

El uso de Cellpose como segmentador generalista, si bien aporta versatilidad y capacidad de adaptación a distintos conjuntos de datos, también introduce una reducción de desempeño en comparación con pipelines diseñados específicamente para un dataset particular con una operación completa, como se observó en la comparación con el estado del arte. En este sentido,

el enfoque adoptado prioriza la generalización por sobre la optimización específica.

Finalmente, si bien la interfaz entrega un grado de explicabilidad visual al mostrar explícitamente las máscaras segmentadas y las células clasificadas como positivas y negativas utilizadas en el cálculo del índice Ki-67, no incorpora mecanismos de explicabilidad interna del modelo. En particular, no se dispone de técnicas que permitan interpretar qué regiones del parche influyen en la decisión del clasificador, como mapas de activación o métodos de atención.

Asimismo, el prototipo no permite la intervención manual del usuario sobre los resultados obtenidos. No es posible corregir segmentaciones erróneas, modificar etiquetas asignadas, ni agregar o eliminar núcleos directamente desde la interfaz. La ausencia de estas herramientas limita su uso como apoyo directo para patólogos y como sistema de anotación asistida para la generación de nuevos conjuntos de datos.

A pesar de estas limitaciones, el prototipo cumple su objetivo como una prueba de concepto funcional. El uso del modelo se justifica en el contexto de un MVP, considerando la escasez de implementaciones públicas que integren segmentación, clasificación nuclear y cuantificación de Ki-67 en una interfaz accesible. Además, la arquitectura propuesta permite, en trabajos futuros, reemplazar o actualizar tanto el segmentador como el clasificador, facilitando la incorporación de modelos entrenados con mejores conjuntos de datos y anotaciones más precisas.

5 Conclusiones y Trabajo Futuro

En esta sección se realiza una síntesis de las fortalezas, debilidades y otros análisis del trabajo completo, así como posibles áreas que mejorar.

5.1 Conclusiones

En este trabajo se propuso e implementó un pipeline completo para la cuantificación automática del índice Ki-67 en imágenes histopatológicas, integrando segmentación nuclear, clasificación basada en parches y una interfaz gráfica funcional orientada a la exploración y análisis de resultados. El enfoque adoptado priorizó la modularidad, la reproducibilidad y la interpretabilidad visual del proceso, por sobre la optimización exclusiva de métricas de desempeño.

Desde el punto de vista metodológico, se demostró que el uso de un segmentador generalista como Cellpose-SAM, combinado con un clasificador entrenado específicamente para distinguir núcleos Ki-67 positivos, negativos y fondo, constituye una aproximación viable para abordar la cuantificación del biomarcador.

Esta estrategia, orientada a usar un segmentador aparte, introduce una dependencia directa al desempeño del mismo. No obstante, también entrega la posibilidad de entrenar al modelo clasificador las salidas esperadas del mismo segmentador en la etapa de inferencia, lo cual resulta clave para la robustez del pipeline completo.

El análisis comparativo entre distintos backbones de clasificación mostró que ConvNeXt-Tiny ofrece el mejor equilibrio global entre desempeño de clasificación nuclear y calidad en la estimación del índice Ki-67 a nivel de imagen. Las diferencias observadas respecto a otras arquitecturas evaluadas fueron acotadas, lo que sugiere que la arquitectura del clasificador no constituye el principal cuello de botella del sistema, sino que el desempeño final está fuertemente influenciado por la etapa de segmentación y por la calidad de los datos disponibles.

En la comparación con trabajos del estado del arte, el pipeline propuesto presentó un desempeño inferior en el conjunto SHIDC-B-Ki-67, mientras que en BCData los resultados fueron comparables, con diferencias moderadas respecto a KPi-Net y mejoras frente a PathoNet en algunas métricas. Este comportamiento refuerza la idea de que la calidad, consistencia y

naturaleza de las anotaciones, así como las características propias de cada dataset, tienen un impacto significativo en el desempeño final, incluso cuando se utilizan arquitecturas y métricas similares.

A nivel de cuantificación del índice Ki-67, se observó que el sistema puede presentar errores locales en la detección o clasificación de núcleos sin que ello se traduzca necesariamente en una degradación proporcional del índice estimado. Métricas como MAE, MSE y la correlación de Pearson mostraron que el pipeline preserva adecuadamente la tendencia global del biomarcador.

La interfaz desarrollada cumple su objetivo como prototipo funcional, permitiendo visualizar de forma explícita las detecciones y clasificaciones que contribuyen al cálculo del índice Ki-67.

Finalmente, la modularidad del sistema y de la interfaz desarrollada constituye uno de los principales aportes de este trabajo, el pipeline permite entrenar distintos modelos de clasificación a partir de anotaciones binarias (positivo y negativo), utilizando máscaras completas generadas por el segmentador. Asimismo, durante la etapa de inferencia, el clasificador puede ser reemplazado o actualizado según el biomarcador o sistema de interés, siempre que el segmentador sea previamente configurado de manera acorde. De esta forma, el prototipo no queda restringido exclusivamente al análisis de Ki-67, sino que se plantea como una base extensible para futuras aplicaciones en patología digital.

Respecto de los objetivos propuestos, se puede afirmar que:

1. Se cumple el desarrollo, a partir de modelos de *Deep Learning*, *Computer Vision* y técnicas afines (utilizando Cellpose y ConvNeXt Tiny), de un sistema para la detección en imágenes de inmunohistoquímica (IHC), específicamente orientado al biomarcador Ki-67.

2. Se cumple la implementación de técnicas de explicabilidad visual para la validación de los resultados de los modelos entrenados, mediante la generación de máscaras, anotaciones de células positivas y negativas sobre la imagen, y un archivo CSV con las coordenadas correspondientes.

3. Se cumple el diseño e implementación de una interfaz que permite cargar imágenes,

procesarlas automáticamente y generar una imagen de salida con los resultados del análisis.

5.2 Trabajo futuro

A partir de los resultados obtenidos, se identifican diversas líneas de trabajo futuro orientadas a mejorar tanto el desempeño como la aplicabilidad del sistema propuesto.

En primer lugar, una extensión natural de este trabajo consiste en incorporar segmentadores especializados o entrenados específicamente sobre los conjuntos de datos utilizados. Si bien el uso de Cellpose permitió una alta versatilidad y una rápida adaptación a distintos datasets, los resultados sugieren que un segmentador ajustado al dominio podría reducir la propagación de errores hacia la etapa de clasificación y mejorar la estimación final del índice Ki-67. En una primera instancia, podría explorarse el ajuste más fino de los parámetros internos de Cellpose, incluso de forma interactiva desde la interfaz. No obstante, una mejora real en la precisión de detección requeriría análisis sistemáticos similares a los realizados en este trabajo, o bien la integración de un segmentador previamente validado para el biomarcador de interés.

De manera similar, otra extensión relevante corresponde a permitir, desde el código fuente, el soporte para múltiples arquitecturas de clasificación. En el prototipo actual, y por motivos de simplicidad, el sistema opera únicamente con ConvNeXt-Tiny y sus respectivos pesos. La incorporación de soporte nativo para distintas arquitecturas permitiría ampliar el espectro de experimentación y adaptar el sistema a distintos escenarios o requerimientos de desempeño.

En relación con la interfaz, si bien esta fue diseñada de forma sencilla y fácilmente modificable a nivel de código, lo cual resulta suficiente para un *Minimum Viable Product* (MVP), una aplicación más madura requeriría ampliar las opciones de configuración accesibles directamente desde el nivel de usuario. En particular, sería deseable permitir la modificación y almacenamiento de parámetros del pipeline sin necesidad de intervención a nivel de desarrollo.

Asimismo, ciertos elementos de la interfaz, como los nombres de las etiquetas y clases, se encuentran actualmente fijados al biomarcador Ki-67, dado que este constituye el único caso de estudio abordado en el trabajo. Como línea de trabajo futuro, se propone generalizar estos componentes para soportar múltiples biomarcadores, permitiendo una configuración flexible

de clases y nomenclaturas desde la interfaz, lo que facilitaría la reutilización del sistema en distintos contextos clínicos y de investigación.

La incorporación de mecanismos de explicabilidad a nivel del clasificador, tales como mapas de activación o técnicas de atención, permitiría analizar con mayor profundidad las decisiones internas del modelo. Si bien la explicabilidad visual basada en máscaras segmentadas y conteo explícito de núcleos resulta más interpretable para el usuario final que un mapa de calor abstracto, este enfoque solo permite observar el resultado del modelo y no los criterios visuales que utiliza para tomar sus decisiones. En este sentido, mecanismos de explicabilidad interna podrían ser útiles para depurar y mejorar el modelo durante su desarrollo.

Otra línea relevante de trabajo futuro corresponde a la integración de herramientas de edición manual dentro de la interfaz, permitiendo al usuario corregir segmentaciones, modificar etiquetas o agregar nuevos núcleos. Esta funcionalidad transformaría la interfaz en una herramienta de anotación asistida, facilitando la creación de nuevos conjuntos de datos y el refinamiento iterativo de los modelos de segmentación y clasificación.

Adicionalmente, un estudio más profundo sobre la influencia de la calidad y consistencia de las anotaciones, idealmente con la participación de patólogos expertos, permitiría evaluar con mayor precisión las causas de las diferencias de desempeño observadas entre los distintos datasets. Este análisis podría extenderse tanto a los conjuntos de entrenamiento como de prueba, contribuyendo a una mejor comprensión de los límites y sesgos del sistema.

Finalmente, el pipeline propuesto fue diseñado de manera modular, lo que abre la posibilidad de extender su uso a otros biomarcadores o tareas de patología digital, reemplazando o ajustando los módulos de segmentación y clasificación según el caso. Permitiendo su evolución a una herramienta general en la detección de biomarcadores.

References

[1] *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 1*. Springer International Publishing, 2020.

[2] Esther Abels, Liron Pantanowitz, Famke Aeffner, Mark D Zarella, Jeroen van der Laak, Marilyn M Bui, Venkata NP Vemuri, Anil V Parwani, Jeff Gibbs, Emmanuel Agosto-Arroyo, Andrew H Beck, and Cleopatra 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.

[3] Hosameldin O. A. Ahmed and Asoke K. Nandi. High performance breast cancer diagnosis from mammograms using mixture of experts with efficientnet features (moeffnet). *IEEE Access*, 12:133703–133725, 2024.

[4] Amir Akbarnejad, Nilanjan Ray, Penny J. Barnes, and Gilbert Bigras. Predicting ki67, er, pr, and her2 statuses from he-stained breast cancer images, 2023.

[5] Cristina Almaraz-López. *Historical Evolution of Deep Learning Applied to Computer Vision*, pages 1–29. 07 2024.

[6] David Anglada-Rotger, Julia Sala, Ferran Marques, Philippe Salembier, and Montse Pardàs. Enhancing ki-67 cell segmentation with dual u-net models: A step towards uncertainty-informed active learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5026–5035, 2024.

[7] Kimberly Ashman, Huimin Zhuge, Erin Shanley, Sharon Fox, Shams Halat, Andrew Sholl, Brian Summa, and J. Quincy Brown. Whole slide image data utilization informed by digital diagnosis patterns. *Journal of Pathology Informatics*, 13:100113, 2022.

[8] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022.

[9] Heang-Ping Chan, Ravi K. Samala, Lubomir M. Hadjiiski, and Chuan Zhou. *Deep Learning in Medical Image Analysis*, pages 3–21. Springer International Publishing, Cham, 2020.

[10] John Chan. The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology. *International Journal of Surgical Pathology*, 22:12–32, 02 2014.

[11] Weimin Chen, Muhammad Ayoub, Mengyun Liao, Ruizheng Shi, Mu Zhang, Feng Su, Zhiguo Huang, Yuanzhe Li, Yi Wang, and Kevin K.L. Wong. A fusion of vgg-16 and vit models for improving bone tumor classification in computed tomography. *Journal of Bone Oncology*, 43:100508, 2023.

[12] Corinna Cortes, Mehryar Mohri, and Yutao Zhong. Improved balanced classification with theoretically grounded loss functions, 2025.

[13] Tejaswini Das, Debasish Swapnesh Kumar Nayak, Anindita Kar, Lambodar Jena, and Tripti Swarnkar. Resnet-50: The deep networks for automated breast cancer classification using mr images. In *2024 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*, pages 1–6, 2024.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[15] Xiaohan Ding, Xiangyu Zhang, Ningning Ma, Jungong Han, Guiguang Ding, and Jian Sun. Repvgg: Making vgg-style convnets great again, 2021.

[16] Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4):198–211, 2007. Computer-aided Diagnosis (CAD) and Image-guided Decision Support.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[18] Savannah R. Duenweg, Samuel A. Bobholz, Allison K. Lowman, Margaret A. Stebbins, Aleksandra Winiarz, Biprojit Nath, Fitzgerald Kyereme, Kenneth A. Iczkowski, and Peter S. LaViolette. Whole slide imaging (wsi) scanner differences influence optical and

computed properties of digitized prostate cancer histology. *Journal of Pathology Informatics*, 14:100321, 2023.

[19] Iván Durán-Díaz, Auxiliadora Sarmiento, Irene Fondón, Clément Bodineau, Mercedes Tomé, and Raúl V. Durán. A robust method for the unsupervised scoring of immunohistochemical staining. *Entropy*, 26(2), 2024.

[20] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[21] Rokshana Stephny Geread, Abishika Sivanandarajah, Emily Brouwer, Geoffrey A. Wood, Dimitrios Androutsos, Hala Faragalla, and April Khademi. pinet: An automated proliferation index calculator framework for ki67 breast cancer images. *bioRxiv*, 2020.

[22] M.L. Giger, N. Karssemeijer, and S.G. Armato. Guest editorial computer-aided diagnosis in medical imaging. *IEEE Transactions on Medical Imaging*, 20(12):1205–1208, 2001.

[23] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, 2019.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[25] Md Shakhawat Hossain, Galib Muhammad Shahriar, M. M. Mahbubul Syeed, Mohammad Faisal Uddin, Mahady Hasan, Shingla Shivam, and Suresh Advani. Region of interest (ROI) selection using vision transformer for automatic analysis using whole slide images. *Scientific Reports*, 13(1):11314, 2023.

[26] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[27] Zhongyi Huang, Yao Ding, Guoli Song, Lin Wang, Ruizhe Geng, Hongliang He, Shan Du, Xia Liu, Yonghong Tian, Yongsheng Liang, S. Kevin Zhou, and Jie Chen. Bcdata:

A large-scale dataset and benchmark for cell detection and counting. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 289–298, Cham, 2020. Springer International Publishing.

[28] Instituto Nacional del Cáncer. Definición de puntuación de Ki-67, 2025. Consultado el 21 de enero de 2026.

[29] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation, 2018.

[30] Kyathanahalli S. Janardhan, Heather Jensen, Natasha P. Clayton, and Ronald A. Herbert. Immunohistochemistry in investigative and toxicologic pathology. *Toxicologic Pathology*, 46(5):488–510, 2018. PMID: 29966501.

[31] Pranav Jeevan and Amit Sethi. Which backbone to use: A resource-efficient domain specific comparison for computer vision, 2025.

[32] Eleanor Jenkinson and Ognjen Arandjelović. Whole slide image understanding in pathology: What is the salient scale of analysis? *BioMedInformatics*, 4(1):489–518, 2024.

[33] Kunal Kawadkar. Comparative analysis of vision transformers and convolutional neural networks for medical image classification, 2025.

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[36] Adam Kukučka, Jan Obdržálek, Vít Musil, Rudolf Nenutil, Petr Holub, and Tomáš Brázdil. Model for ki-67 proliferation index predicton, trained end-to-end on routine diagnostic data. *medRxiv*, 2025.

[37] Zhiqiang Lao, Dinggang Shen, Dengfeng Liu, Abbas F. Jawad, Elias R. Melhem, Lenore J. Launer, R. Nick Bryan, and Christos Davatzikos. Computer-assisted segmentation of white matter lesions in 3d mr images using support vector machine. *Academic Radiology*, 15(3):300–313, 2008.

[38] M. Latha, P. Santhosh Kumar, R. Roopa Chandrika, T. R. Mahesh, V. Vinoth Kumar, and Suresh Guluwadi. Revolutionizing breast ultrasound diagnostics with efficientnet-b7 and explainable ai. *BMC Medical Imaging*, 24(1):230, 2024.

[39] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[40] Lian-Tao Li, Guan Jiang, Qian Chen, and Jun-Nian Zheng. Ki67 is a promising molecular target in the diagnosis of cancer (review). *Molecular Medicine Reports*, 11(3):1566–1572, 2015.

[41] Sylwia Libard, Dijana Cerjan, and Irina Alafuzoff. Characteristics of the tissue section that influence the staining outcome in immunohistochemistry. *Histochemistry and Cell Biology*, 151(1):91–96, 2019.

[42] Qi Liu, Zhenfeng Zhao, Lei Lou, Yuehong Li, and Shenwen Wang. Multi scale deep learning quantifies ki67 index in breast cancer histopathology images. *Scientific Reports*, 15(1):44972, 2025.

[43] Yiqing Liu, Xi Li, Aiping Zheng, Xihan Zhu, Shuting Liu, Mengying Hu, Qianjiang Luo, Huina Liao, Mubiao Liu, Yonghong He, and Yupeng Chen. Predict ki-67 positive cells in he-stained images using deep learning independently from ihc-stained images. *Frontiers in Molecular Biosciences*, Volume 7 - 2020, 2020.

[44] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022.

[45] Markus Marks, Uriah Israel, Rohit Dilip, Qilin Li, Changhua Yu, Emily Laubscher, Ahamed Iqbal, Elora Pradhan, Ada Ates, Martin Abt, Caitlin Brown, Edward Pao, Shenyi Li, Alexander Pearson-Goulart, Pietro Perona, Georgia Gkioxari, Ross Barnowski, Yisong Yue, and David Van Valen. Cellsam: a foundation model for cell segmentation. *Nature Methods*, 22(12):2585–2593, 2025.

[46] Jamie D. Martina, Christopher Simmons, and Drazen M. Jukic. High-definition hematoxylin and eosin staining in a transition to digital pathology. *Journal of Pathology Informatics*, 2(1):45, 2011.

[47] National Cancer Institute. Tests for breast cancer biomarkers, dec 2025. Consultado el 21 de enero de 2026.

[48] Farzin Negahbani, Rasool Sabzi, Bita Pakniyat Jahromi, Dena Firouzabadi, Fateme Movahedi, Mahsa Kohandel Shirazi, Shayan Majidi, and Amirreza Dehghanian. Pathonet introduced as a deep neural network backend for evaluation of ki-67 and tumor-infiltrating lymphocytes in breast cancer. *Scientific Reports*, 11(1):8489, 2021.

[49] Liron Pantanowitz, Ashish Sharma, Alexis B. Carter, Tahsin Kurc, Alan Sussman, and Joel Saltz. Twenty years of digital pathology: An overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *Journal of Pathology Informatics*, 9(1):40, 2018.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[51] Lawrence G. Roberts. *Machine perception of three-dimensional solids.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1963. Thesis (Ph. D.)–Massachusetts Institute of Technology, Dept. of Electrical Engineering.

[52] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[53] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025.

[54] Caglar Senaras, M Khalid Khan Niazi, Gerard Lozanski, and Metin N Gurcan. Deepfocus: detection of out-of-focus regions in whole slide digital images using deep learning. *PloS one*, 13(10):e0205387, 2018.

[55] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[56] Fabio A. Spanhol, Luiz S. Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7):1455–1462, 2016.

[57] Carsen Stringer, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *bioRxiv*, 2020.

[58] Carsen Stringer and Marius Pachitariu. Cellpose3: one-click image restoration for improved cellular segmentation. *Nature Methods*, 22(3):592–599, 2025.

[59] Md. Alamin Talukder. An improved xai-based densenet model for breast cancer detection using reconstruction and fine-tuning. *Results in Engineering*, 26:104802, 2025.

[60] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

[61] Mingxing Tan and Quoc V. Le. Efficientnetv2: Smaller models and faster training, 2021.

[62] Gei Ki Tang, Chee Chin Lim, Faezahtul Arbaeyah Hussain, Qi Wei Oung, Aidy Irman Yajid, Sumayyah Mohammad Azmi, and Yen Fook Chong. Enhanced hovernet optimization for precise nuclei segmentation in diffuse large b-cell lymphoma. *Diagnostics*, 15(15), 2025.

[63] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention, 2021.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[65] Zhen Wang, Shuang Fu, Hongguang Zhang, Chunyang Wang, Chunhui Xia, Pen Hou, Chunxue Shun, and Ge Shun. Dual-branch dynamic hierarchical u-net with multi-layer space fusion attention for medical image segmentation. *Scientific Reports*, 15(1):8194, 2025.

[66] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text, 2022.

[67] Martin Weigert and Uwe Schmidt. Nuclei instance segmentation and classification in histopathology images with stardist. In *2022 IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*, page 1–4. IEEE, March 2022.

[68] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, Kun-Hsing Yu, Sierra Willens, Francesca Maria Olguin, Jeffrey J. Nirschl, Joel Neal, Maximilian Diehn, Sen Yang,

and Ruijiang Li. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

[69] Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu, and Jiang Bian. Me llama: Foundation large language models for medical applications, 2024.

[70] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks, 2017.

[71] Chhavi Yadav and Léon Bottou. Cold case: The lost MNIST digits. *CoRR*, abs/1905.10498, 2019.

[72] Pengyuan Zhang, Weihao Jin, Jiahong Li, and Jiao Tian. Application of improved hover-net in nuclear segmentation and classification of histopathological images. In *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA)*, pages 351–357, 2024.

Apéndice

Los códigos y *scripts* empleados en el desarrollo de este trabajo, así como los entornos, su configuracion, prototipo de interfaz y pesos se encuentran en el repositorio asociado: https://github.com/Pokex-LL/biomarker-ki67