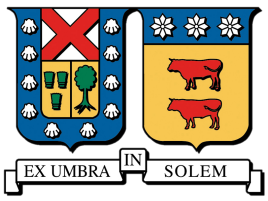


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA

Tesis de Magíster

para la obtención del grado académico de
Magíster en Ciencias de la Ingeniería Informática

Explicabilidad Visual para Diagnóstico Histopatológico: Generando Contrafactuales a través de GANs

Branndon Felipe Muñoz Pinto

Directora de Tesis: Dr. Raquel Pezoa Rivera Universidad Técnica Federico Santa María, Chile

2025

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: Visual Explanations on Histopathology Image Classification: Generating Counterfactuals through GANs

Nombre del candidato(a): Branndon Felipe Muñoz Pinto

Carrera / Grado: Magíster en ciencias de la Ingeniería en Informática

Campus: Casa Central Valparaiso **;** **Departamento**: Informática

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Raquel Pezoa Rivera, 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

☒ 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:

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

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):
Se publicará un segundo paper de los resultados finales del trabajo y hasta entonces no se pueden dar a conocer los detalles de antemano.

4.- FIRMAS

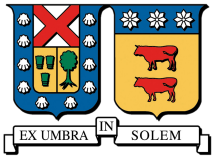
Profesor(a) guía o director(a) de memoria o tesis:

Fecha:29/07/2025; **Firma:** _____

Estudiante o Candidato(a):

Fecha: 24/07/2025 ; **Firma:** _____

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.



TÍTULO DE LA TESIS:

EXPLICABILIDAD VISUAL PARA DIAGNÓSTICO HISTOPATOLÓGICO: GENERANDO CONTRAFACTUALES A TRAVÉS DE GANS

AUTOR:

BRANNDON FELIPE MUÑOZ PINTO

Tesis presentada como requerimiento para optar al grado de
Magíster en Ciencias de la Ingeniería Informática de la Universidad Técnica Federico Santa María.

Directora de Tesis:

Dr. Raquel Pezoa Rivera
Departamento de Informática
Universidad Técnica Federico Santa María

Profesor Correferente:

Dr. Julio Sotelo
Departamento de Informática
Universidad Técnica Federico Santa María

Examinador Externo Nacional:

Dr. Rodrigo Salas
Escuela de Ingeniería Biomédica
Universidad de Valparaíso

Declaración de autoría

Por la presente declaro que he redactado esta tesis sobre el tema

Explicabilidad Visual para Diagnóstico Histopatológico: Generando Contrafactuales a través de GANs

de forma independiente. No utilicé más ayudas, fuentes, figuras ni recursos que los indicados en las referencias. Todas las partes tomadas de otras fuentes están claramente marcadas y citadas correctamente.

Asimismo, declaro que —según mi leal saber y entender— este trabajo o partes del mismo no han sido presentados anteriormente por mí ni por otra persona en esta ni en ninguna otra universidad.

Branndon Felipe Muñoz Pinto

Valparaíso, August 29, 2025 _____

Agradecimientos

Quiero expresar mi más profundo agradecimiento a mi familia, por ser siempre mi pilar incondicional. Gracias por brindarme su apoyo constante, por preocuparse de que mi única preocupación fuera concentrarme en mis estudios, y por estar presentes en cada etapa de este camino.

A mi polola, gracias por tu amor, por contenerme en los momentos difíciles y por hacer siempre todo lo posible para mantener viva mi motivación. Gracias por tu paciencia, tu compañía y por estar a mi lado incluso cuando no estuve en mi mejor versión.

A mi hermano, gracias por enseñarme el valor del temple, por tu ejemplo y por siempre llegar justo cuando más necesitaba una palabra de aliento o una muestra de cariño.

A mis abuelos, tíos y primos, les agradezco profundamente por su preocupación, por desearme siempre lo mejor y por abrirme las puertas de su hogar durante los veranos en el sur, con toda la paciencia y cariño que eso conlleva.

A mis amigos y a todas aquellas personas que, con una risa, un consejo, un momento o un pequeño gesto, contribuyeron a mi crecimiento personal y académico. Cada uno aportó con un granito de arena a esta etapa tan significativa de mi vida.

Agradezco especialmente a mi profesora Raquel por su gran apoyo, por su dedicación, su forma de enseñar y motivar, y por haber estado siempre dispuesta a ayudar en todo lo necesario para sacar adelante esta tesis. También a Helen, por entregarme su tiempo, apoyo y conocimiento experto en el dominio, lo cual fue fundamental para avanzar en esta investigación.

Finalmente, agradezco a mi Universidad por la formación recibida y el apoyo económico otor-

gado a través de las becas de Arancel y de Mantención. Este trabajo fue posible gracias al uso de de los créditos de AWS-U.Chile-NLHPC. Powered@NLHPC: Esta investigación fue parcialmente apoyada por el National Laboratory for High Performance Computing (NLHPC) de la Universidad de Chile. Se agradece también el apoyo parcial del proyecto ANID PIA/APOYO AFB230003.

Resumen

La creciente integración de la inteligencia artificial (IA) en el diagnóstico médico, particularmente en el análisis de imágenes histopatológicas, subraya la necesidad crítica de modelos que no solo sean precisos, sino también interpretables. Esta tesis aborda el desafío de la explicabilidad en el aprendizaje profundo mediante el desarrollo y la evaluación de un sistema avanzado para la generación de imágenes histopatológicas sintéticas y la creación de explicaciones contrafactuales visuales. El sistema se fundamenta en una arquitectura StyleGAN2-ADA, complementada con un codificador para la proyección de imágenes reales al espacio latente y un clasificador para guiar la semántica de las imágenes generadas.

Se investigó exhaustivamente la optimización del proceso de entrenamiento, demostrando que una estrategia en dos etapas —priorizando la reconstrucción antes de introducir la tarea de clasificación— mejora significativamente la calidad de las imágenes generadas, logrando un Fréchet Inception Distance (FID) de 16,2. Asimismo, se analizó el impacto de diferentes formulaciones de la función de pérdida, identificando que una configuración balanceada (Learned Perceptual Image Patch Similarity (LPIPS) de 0,15, Peak Signal-to-Noise Ratio (PSNR) de 28,0 dB y Structural Similarity Index Measure (SSIM) de 0,80) ofrece un equilibrio óptimo entre fidelidad numérica, calidad perceptual y consistencia en el espacio latente. La exploración de este espacio latente reveló su capacidad para organizar las características histopatológicas de manera desentrelazada y semánticamente coherente, diferenciando claramente entre tejido benigno y canceroso y capturando la variabilidad entre clases.

El núcleo de este trabajo radica en un novedoso método para generar contrafactuales visuales, que ilustran cómo modificaciones mínimas y específicas en una imagen pueden alterar la predicción de un clasificador. Se demostró la capacidad de generar trayectorias de interpolación latente que visualizan la transición gradual entre clases, identificando puntos de inflexión en la decisión del clasificador. Además, se comparó el método propuesto con “Chexplaining in Style”, demos-

trando una eficiencia computacional drásticamente superior (3,1s frente a 224,7s para generar 50 contrafactuales) y una mayor aplicabilidad al dominio histopatológico, al generar explicaciones diversas y semánticamente coherentes, a diferencia del método de referencia que no logró inducir cambios de clase relevantes.

La validación del sistema incluyó una evaluación cuantitativa y cualitativa, así como una extensa evaluación interdisciplinaria con 13 profesionales de la salud, incluyendo tecnólogos médicos, bioquímicos y médicos con un promedio de 11 años de experiencia. Los resultados indicaron un alto grado de realismo en las imágenes sintéticas (70% de fool rate) , una excelente conservación de patrones biológicos en las reconstrucciones (realismo visual promedio de 4,12 sobre un máximo de 5) y una alta utilidad percibida de los contrafactuales como herramienta de apoyo diagnóstico (92,3% de calificación “muy útil” o “extremadamente útil”), especialmente en casos complejos.

Este trabajo contribuye significativamente al campo de la IA explicable (xAI) en histopatología, ofreciendo un marco robusto para generar imágenes de alta calidad y explicaciones visuales intuitivas que pueden mejorar la confianza y la comprensión de los modelos de aprendizaje profundo en aplicaciones clínicas críticas. Las metodologías y hallazgos presentados sientan las bases para futuras investigaciones orientadas a la integración de estas herramientas en la práctica diagnóstica.

Abstract

The increasing integration of artificial intelligence (AI) into medical diagnosis, particularly in the analysis of histopathological images, underscores the critical need for models that are not only accurate but also interpretable. This thesis addresses the challenge of explainability in deep learning by developing and evaluating an advanced system for the generation of synthetic histopathological images and the creation of visual counterfactual explanations. The system is based on a StyleGAN2-ADA architecture, complemented by an encoder for projecting real

images into the latent space and a classifier to guide the semantics of the generated images.

The training process was thoroughly optimized, demonstrating that a two-stage strategy—prioritizing reconstruction before introducing the classification task—significantly improves the quality of the generated images, achieving an Fréchet Inception Distance (FID) of 16.2. Furthermore, the impact of different loss formulations was analyzed, identifying that a balanced configuration (Learned Perceptual Image Patch Similarity (LPIPS) of 0,15, Peak Signal-to-Noise Ratio (PSNR) of 28,0 dB and Structural Similarity Index Measure (SSIM) of 0,80) offers an optimal trade-off between numerical fidelity, perceptual quality, and consistency in the latent space. Exploration of this latent space revealed its ability to organize histopathological features in a disentangled and semantically coherent manner, clearly distinguishing between benign and cancerous tissue while capturing intra-class variability.

The core of this work lies in a novel method for generating visual counterfactuals, which illustrate how minimal, specific modifications to an image can alter a classifier's prediction. The method enables the generation of latent interpolation trajectories that visualize the gradual transition between classes, identifying decision boundary inflection points. Moreover, our proposed method was compared to "Chexplaining in Style", demonstrating drastically superior computational efficiency (3.1s vs. 224.7s to generate 50 counterfactuals) and greater applicability to the histopathology domain by producing diverse and semantically coherent explanations, unlike the reference method which failed to induce relevant class changes.

The system's validation included quantitative and qualitative analysis, as well as an extensive interdisciplinary evaluation with 13 healthcare professionals, including medical technologists, biochemists, and physicians with an average of 11 years of experience. Results showed a high degree of realism in synthetic images (70% fool rate) , excellent preservation of biological patterns in reconstructions (average visual realism of 4.12/5) , and high perceived usefulness of counterfactuals as a diagnostic support tool (92.3% rated "very useful" or "extremely useful"), especially in complex cases.

This work makes a significant contribution to the field of explainable AI (XAI) in histopathology, providing a robust framework for generating high-fidelity images and intuitive visual ex-

planations that can improve trust and understanding of deep learning models in critical clinical applications. The methodologies and findings presented lay the groundwork for future research aimed at integrating these tools into diagnostic practice.

Contents

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

1.1. Contexo del Problema

La histopatología es el estudio de los tejidos biológicos a nivel microscópico, esencial para el diagnóstico de una variedad de enfermedades, incluyendo el cáncer. Tradicionalmente, este proceso implica la observación de muestras teñidas a través de un microscopio, donde los patólogos evalúan características morfológicas específicas para identificar anormalidades. Sin embargo, el campo ha evolucionado considerablemente con el advenimiento de la histopatología digital, que permite la digitalización de imágenes histopatológicas de diapositiva completas en alta resolución, conocidas como Whole Slide Images (WSI) (ver Figura 1.1). Este avance permite almacenar, analizar y compartir imágenes patológicas a gran escala y ha revolucionado la manera en que se diagnostican las enfermedades [Madabhushi and Lee, 2016].

1.1.1. Obtención de Imágenes Histopatológicas Digitales

El proceso de creación de una imagen digital de tejido histopatológico comienza con la preparación y tinción de la muestra. Comúnmente, los tejidos son tratados con hematoxilina y eosina (H&E), un método de tinción que permite resaltar estructuras celulares clave. Luego, se escanean las láminas usando escáneres de alta precisión que pueden capturar detalles minuciosos de

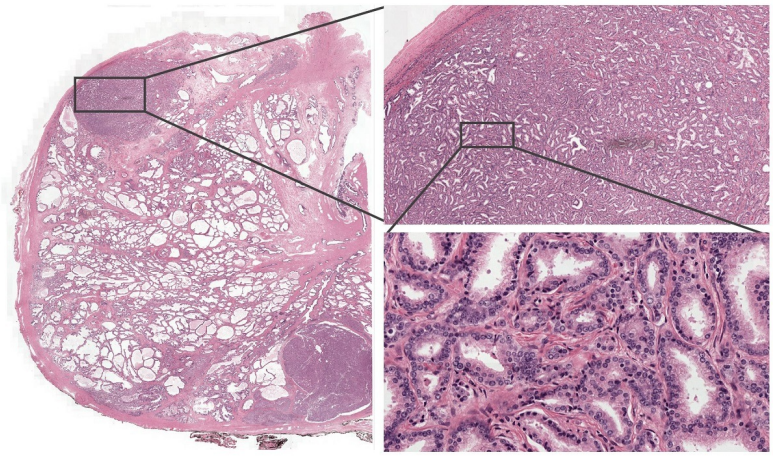


Figura 1.1: Ejemplo de imagen histopatológica de diapositiva completa (WSI) del dataset PESO [Bulten et al., 2019].

la arquitectura tisular en niveles de magnificación elevados. Las imágenes obtenidas son de alta resolución y pueden llegar a tener tamaños de varios gigabytes, lo cual permite a los patólogos observar los tejidos a diferentes niveles de aumento, replicando de manera digital la experiencia del microscopio tradicional [Bera et al., 2019].

Las imágenes digitales así generadas pueden almacenarse en sistemas de gestión de datos y ser accesibles para el análisis computacional. Este formato digital permite el uso de algoritmos de aprendizaje profundo para la detección automática de patrones, cuantificación de células, y clasificación de anormalidades [Ignatov et al., 2024]. La capacidad de analizar de forma computarizada estas imágenes en busca de características sutiles en grandes volúmenes de datos ha llevado a una creciente adopción de la histopatología digital en el entorno clínico.

1.1.2. Desafíos en el Análisis de Imágenes Histopatológicas

A pesar de los avances en la digitalización de imágenes histológicas, existen desafíos significativos. La alta resolución de las imágenes y la complejidad de las características morfológicas plantean importantes desafíos para la clasificación precisa de tejidos. Esto se debe a la gran variabilidad en la forma y textura de las estructuras que los conforman, como células, estroma y

linfocitos, entre otros. Asimismo, la escasez de datos disponibles dificulta aún más este proceso, ya que los conjuntos de datos suelen estar categorizados según la zona de extracción del tejido. Por ejemplo, un tejido canceroso de colon no necesariamente presenta el mismo aspecto ni los mismos componentes que uno de mama. Además, factores como la calidad de la tinción y las diferencias en los protocolos de preparación pueden alterar la apariencia visual de las muestras, complicando aún más el análisis [Komura and Ishikawa, 2018].

Los modelos de aprendizaje profundo que se emplean en la clasificación de estas imágenes presentan un desafío crítico y poco conocido: la interpretabilidad. Los algoritmos de caja negra, aunque efectivos, no ofrecen una comprensión clara de los factores específicos que influyen en una predicción. Sin embargo, la necesidad de contar con modelos de clasificación en este contexto es evidente. Las imágenes histopatológicas contienen una enorme cantidad de información visual, cuyo análisis manual requiere un tiempo considerable y está sujeto a variabilidad inter-observador. Un modelo de clasificación permite automatizar la identificación de patrones histológicos relevantes, reduciendo la carga de trabajo de los especialistas y agilizando el proceso diagnóstico. En el contexto clínico, los patólogos y médicos requieren explicaciones interpretables para **validar las decisiones del modelo y asegurar que los resultados son clínicamente significativos y libres de sesgos** [Lundervold and Lundervold, 2019]. Este aspecto es particularmente importante cuando se utilizan algoritmos de IA para el diagnóstico de enfermedades graves, ya que cualquier error o sesgo puede tener consecuencias críticas para el paciente.

1.1.3. Importancia de las Explicaciones Contrafactuales en Histopatología Digital

En este contexto, existen numerosos e innovadores métodos que permiten hacer a los modelos de IA más interpretables, conocidos como métodos de Explicabilidad de Inteligencia Artificial (xAI). Sin embargo, no todos pueden ofrecer el mismo valor para el contexto de histopatología en particular. Según un estudio de encuesta a múltiples patólogos [Evans et al., 2022], el razonamiento contrafactual habría recibido la respuesta más positiva en general comparado a los otros métodos de explicabilidad, haciendo énfasis en que este tipo de razonamiento ayuda a entender

lo que el algoritmo está buscando. Además, existen múltiples trabajos que validan su uso en entornos clínicos y diagnóstico médico [Tanyel et al., 2023].

No obstante, la implementación de un método contrafactual en imágenes no es trivial. Los métodos contrafactuales tradicionales a menudo fallan en generar resultados plausibles. Estos métodos tienden a modificar píxeles o superpíxeles específicos que dan lugar a una imagen que no existe en la realidad o que no es coherente con el contexto original [Nemirovsky et al., 2021]. Para intentar resolver esto han surgido algunos métodos de transferencia de estilo, pero estos solo se limitan a manipular aspectos como el color o textura y no pueden alterar la forma o estructura de las imágenes de manera significativa, lo que es crucial para una interpretación profunda y precisa en el contexto histopatológico y médico en general.

Para abordar todas estas problemáticas, esta tesis se enmarca en el contexto de Explainable Artificial Intelligence (xAI) y propone un nuevo método de explicabilidad basado en Redes Generativas Adversarias (GAN) y métodos contrafactuales. La aplicación de este nuevo método será en el área de histopatología, más precisamente se explicarán modelos de redes neuronales de clasificación de imágenes histopatológicas. En particular se utilizará el enfoque StyleGAN2 ADA ya que permite generar imágenes de alta calidad y se adapta al contexto histopatológico, donde a menudo se dispone de datos limitados. Además, este método hace posible generar una representación desentrelazada del espacio latente, pudiendo así manipular de manera precisa las características latentes. Utilizando estas representaciones en conjunto con un clasificador entrenado sobre el mismo conjunto de datos de imágenes histopatológicas, se diseñará un método para generar explicaciones contrafactuales con la premisa de proporcionar un mayor nivel de interpretabilidad en las decisiones de clasificación. El impacto de estas explicaciones en la confianza y la interpretabilidad del modelo será evaluado mediante estudios cuantitativos y cualitativos, integrando retroalimentación de expertos del área histopatológica y análisis detallado de los resultados obtenidos.

1.2. Hipótesis

- Hipótesis 1: La implementación de un modelo generativo basado en StyleGAN2 ADA sobre un dataset de imágenes histopatológicas específico permitirá generar imágenes realistas y representaciones desentrelazadas en el espacio latente.
- Hipótesis 2: La generación de explicaciones contrafactuales mediante la manipulación del espacio latente mejorará la interpretabilidad y la confianza en las decisiones de clasificación de imágenes histopatológicas.

1.3. Objetivos

1.3.1. Objetivos Generales

Diseñar, implementar y evaluar un método para generar explicaciones contrafactuales interpretables en imágenes histopatológicas mediante la manipulación del espacio latente desentrelazado obtenido a partir del entrenamiento de un modelo basado en Redes Generativas Adversarias, en particular, StyleGAN2 ADA, además de un clasificador y un encoder.

1.3.2. Objetivos Específicos

- Desarrollar y entrenar un modelo generativo basado en StyleGAN2 ADA para la generación de imágenes histopatológicas realistas con un espacio latente desentrelazado.
- Desarrollar e implementar un método para generar explicaciones contrafactuales basado en la manipulación del espacio latente del modelo generativo.
- Evaluar cuantitativa y cualitativamente los resultados para medir el realismo de las imágenes generadas y la interpretabilidad del clasificador.

1.4. Estructura

El presente documento está estructurado de la siguiente manera: en el Capítulo 2 se presentará el Marco Teórico, donde se abordarán los conceptos fundamentales relacionados con la interpretabilidad en inteligencia artificial, diversos tipos de métodos de explicabilidad, GANs, representaciones latentes y métodos de evaluación. En el Capítulo 3 se realizará una revisión del estado del arte, analizando los trabajos más relevantes que han intentado abordar la falta de interpretabilidad en modelos generativos aplicados a la medicina. Además, se describirá la Metodología empleada, detallando el diseño experimental y los pasos para el desarrollo del modelo. En el Capítulo 4 se analizarán los resultados obtenidos, comparándolos con trabajos previos. Finalmente, en el Capítulo 5 se presentarán las conclusiones del trabajo y se propondrán posibles líneas de investigación futura.

Capítulo 2

Marco Teórico

2.1. La importancia de IA explicable (xAI)

La inteligencia artificial (IA) se está integrando cada vez más en sectores críticos como la medicina, las finanzas, la justicia y el transporte, donde sus decisiones tienen un impacto directo en la vida de las personas. Esta creciente aplicación hace que la interpretabilidad y explicabilidad de los modelos de IA sean fundamentales. Los modelos de IA actuales, especialmente aquellos basados en redes neuronales profundas, suelen funcionar como una "caja negra". Esto significa que, aunque sean capaces de ofrecer predicciones precisas, resulta difícil entender cómo se llegó a una determinada decisión. La falta de transparencia impide que estas tecnologías sean aceptadas, especialmente en contextos donde la confianza del usuario es fundamental. En sectores como la medicina y el sistema judicial, las decisiones automatizadas requieren justificación y validación, ya que la ausencia de explicaciones claras genera incertidumbre sobre sus mecanismos, lo que genera desconfianza y limita su uso en la práctica.

El avance de la IA en áreas críticas como la patología digital está impulsado por la proliferación de flujos de trabajo digitalizados y los avances en el aprendizaje automático (ML). Estos desarrollos han permitido optimizar la atención a los pacientes y reducir la carga de trabajo de los profesionales, pero al mismo tiempo, han puesto de manifiesto las limitaciones inherentes a

los modelos de caja negra, en los cuales la falta de interpretabilidad representa una barrera significativa para la aceptación y el uso clínico seguro. En este contexto, la IA explicable (xAI) se presenta como una respuesta clave para mitigar esta problemática, ya que proporciona un medio para entender las decisiones del modelo y facilita una mejor integración en los flujos de trabajo existentes.

Las consecuencias de utilizar modelos no explicables en contextos sensibles pueden ser desastrosas. Existen ejemplos documentados de sistemas de IA que han cometido errores graves, como diagnósticos médicos incorrectos o decisiones basadas en sesgos, precisamente debido a la incapacidad de explicar sus predicciones [Dastin, 2018, Obermeyer et al., 2019]. Estos incidentes resaltan la importancia de disponer de mecanismos que permitan entender y evaluar el comportamiento de los modelos de IA, para prevenir errores que puedan tener consecuencias severas. Además, el panorama normativo y las regulaciones en desarrollo exigen una mayor transparencia en el uso de IA en medicina, lo cual obliga a la comunidad científica a concentrarse en la creación de modelos más interpretables y que puedan cumplir con estos estándares de seguridad y ética.

La IA explicable (xAI) ofrece una serie de ventajas significativas. Permite la identificación y corrección de sesgos presentes en los modelos, contribuyendo a desarrollar sistemas más justos y éticos. Además, la capacidad de explicar cómo un modelo llega a sus conclusiones es esencial para validar y confiar en sus recomendaciones. En el ámbito médico, los profesionales de la salud no solo requieren predicciones precisas, sino también comprender las razones subyacentes que sustentan dichas predicciones. La capacidad de justificar una decisión médica automatizada puede marcar la diferencia entre un diagnóstico confiable y una recomendación que los médicos decidan ignorar.

A pesar de sus evidentes beneficios, implementar xAI presenta desafíos importantes. Los investigadores se enfrentan a la complejidad de hacer que los modelos sean explicables sin sacrificar su precisión. En muchos casos, existe un compromiso entre el rendimiento del modelo y su capacidad de ser interpretado, lo cual dificulta el desarrollo de soluciones xAI que sean eficientes y fiables.

2.2. Perspectiva general de explicabilidad en Inteligencia Artificial

La explicabilidad en inteligencia artificial (xIA) se refiere a la capacidad de los modelos para proporcionar razones comprensibles sobre cómo y por qué se llegó a una determinada decisión. En el contexto de xAI, no existe una categorización como tal, pero según algunos autores [Doshi-Velez and Kim, 2017, Lipton, 2018] los modelos de IA se pueden clasificar en dos categorías principales: los que son inherentemente explicables, como los árboles de decisión y los modelos lineales, y los llamados "modelos de caja negra", como las redes neuronales profundas. Estos últimos, aunque presentan mayor precisión, carecen de interpretabilidad. En la Figura 2.1 se ilustra de mejor manera la relación que existe entre la interpretabilidad y la precisión en distintos modelos de machine learning.

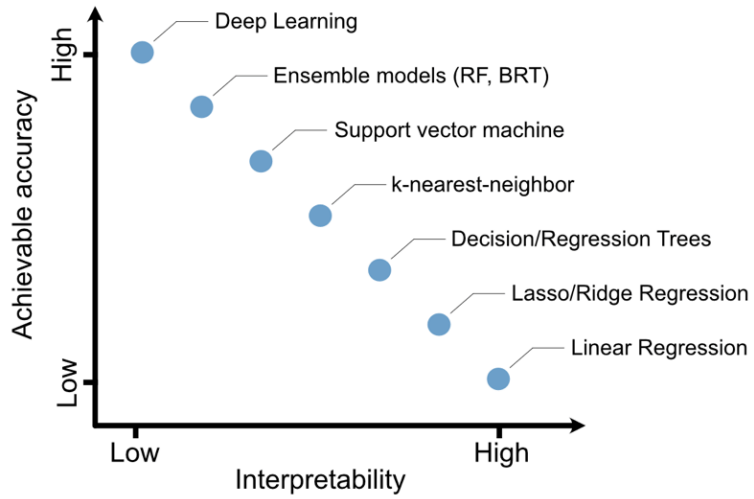


Figura 2.1: **Relación entre interpretabilidad y precisión en distintos modelos de machine learning [Pichler and Hartig, 2022]**. A medida que aumenta la precisión (eje vertical), la interpretabilidad del modelo (eje horizontal) tiende a disminuir. Los modelos más simples, como la regresión lineal, son altamente interpretables pero menos precisos, mientras que los modelos complejos como los de deep learning ofrecen mayor precisión a costa de menor interpretabilidad.

Sin una explicación clara, los sistemas de IA pueden ser percibidos como poco fiables. Este dilema ha dado lugar a un campo de investigación enfocado en el desarrollo de técnicas de explicabilidad que permitan superar las barreras inherentes a los modelos de caja negra, promoviendo así una adopción más segura y efectiva de la IA.

2.2.1. Definiciones y Problemáticas

En la literatura, los términos explicabilidad e interpretabilidad no siempre tienen definiciones consistentes, y aunque existen autores que los usan indistiguidamente, en el trabajo [Gilpin, 2018] hacen énfasis en que hay razones suficientes para diferenciarlos:

- **Explicabilidad:** Describe cómo funciona un modelo y los mecanismos involucrados en generar una predicción específica o un tipo de predicción. Se enfoca en desglosar los detalles del proceso de decisión del modelo, ofreciendo una descripción precisa de las operaciones internas. La explicabilidad busca hacer transparentes los componentes internos de los modelos, proporcionando información detallada sobre cómo se generan los resultados.
- **Interpretabilidad:** Se refiere a la capacidad de describir el sistema de una manera comprensible para los usuarios finales, sin necesidad de analizar los detalles que llevan a la decisión del modelo. Está más relacionada con el contexto de aplicación, así como con el conocimiento y los sesgos de los usuarios. La interpretabilidad tiene como objetivo facilitar la comprensión del comportamiento del modelo a nivel general, de modo que los usuarios puedan entender las predicciones sin profundizar en la complejidad del proceso subyacente.

El término “explicabilidad de inteligencia artificial” se utiliza para describir la capacidad de un sistema de IA para proporcionar descripciones que permitan a los usuarios humanos entender el proceso detrás de sus decisiones. Este concepto está estrechamente relacionado con la interpretabilidad, que implica el grado en el que un humano puede entender la causa detrás de una predicción específica. La principal problemática de los modelos de caja negra radica en que, aunque logran altos niveles de precisión, no proporcionan una justificación comprensible de

sus decisiones. Esto hace que, en aplicaciones críticas, sea difícil para los usuarios aceptar el resultado del modelo sin reservas.

Por ejemplo, en el ámbito de la medicina, un modelo de clasificación de imágenes puede detectar una anomalía con gran precisión, pero si no puede explicar qué características de la imagen llevaron a esa conclusión, el diagnóstico resultante podría ser cuestionado por los médicos. Además, los modelos de aprendizaje profundo pueden aprender correlaciones engañosas, es decir, patrones en los datos que no son relevantes para el problema real pero que se confunden con características significativas, lo cual puede conducir a decisiones incorrectas. Estas limitaciones hacen necesario el desarrollo de métodos que permitan que los modelos sean tanto precisos como explicables.

2.2.2. Tipos de explicabilidad

Existen diferentes enfoques para proporcionar explicabilidad en sistemas de IA, que se pueden clasificar principalmente en dos tipos: la explicabilidad intrínseca y la explicabilidad post hoc [Lipton, 2018, Doshi-Velez and Kim, 2017].

- **Explicabilidad Intrínseca:** Este tipo de explicabilidad se refiere a la capacidad de algunos modelos para ser comprensibles por sí mismos. Modelos como los árboles de decisión, las regresiones lineales y las reglas basadas en lógica son ejemplos de modelos intrínsecamente explicables. Estos modelos se diseñan de tal forma que sus procesos de toma de decisiones sean fácilmente entendibles para los humanos, lo que hace que sean ideales para aplicaciones donde la transparencia es una prioridad.
- **Explicabilidad Post-Hoc:** En contraste, la mayoría de los modelos de caja negra no son explicables de manera intrínseca, y requieren técnicas adicionales para proporcionar explicaciones una vez que se ha generado una predicción. La explicabilidad post hoc puede incluir el uso de saliency maps, contrafactuales, entre otros. Estas técnicas buscan traducir las decisiones del modelo en términos que sean comprensibles para los usuarios, facilitando la comprensión del comportamiento del sistema.

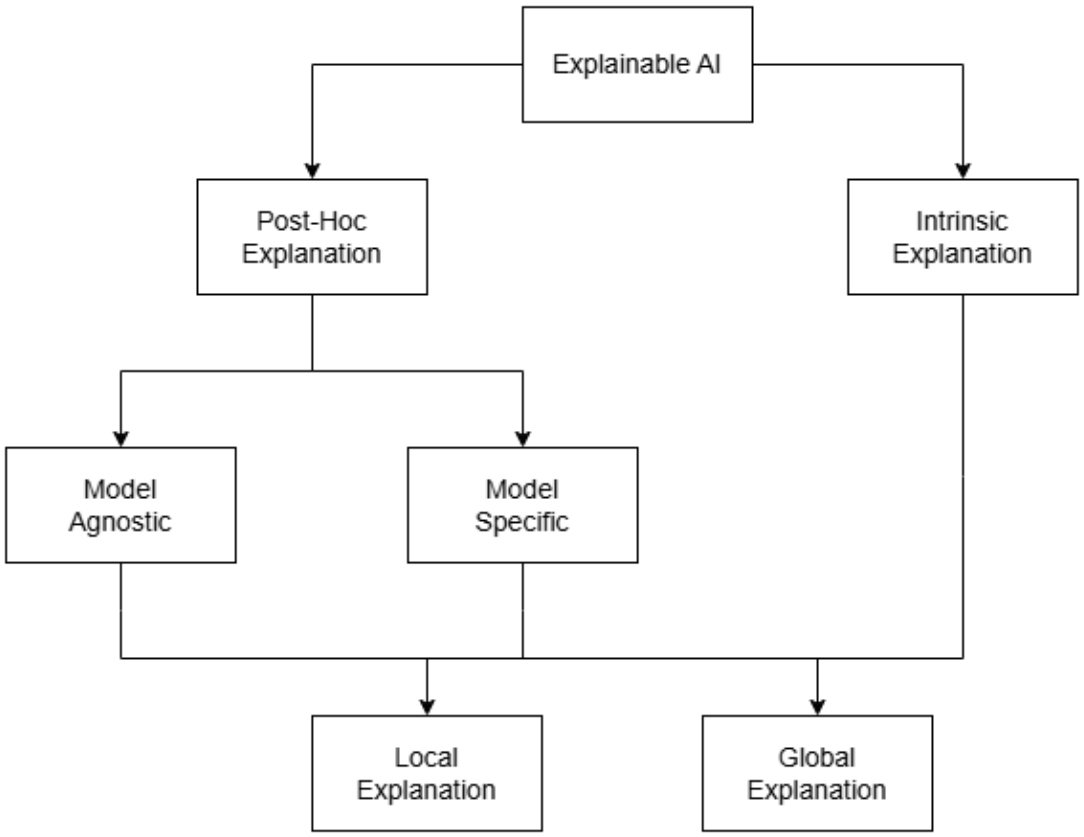


Figura 2.2: Taxonomía de los métodos de explicabilidad.

2.3. Explicaciones Contrafactuales Visuales

Los métodos contrafactuales visuales buscan responder preguntas del tipo: ¿qué cambios serían necesarios en la imagen de entrada para que el modelo la clasificara como una clase diferente c_{counter} en lugar de la clase original c? Para ello, generan ejemplos que pertenecen a la distribución real de datos, es decir, imágenes visualmente plausibles que podrían haber sido observadas durante el entrenamiento, y que ilustran explícitamente cómo debería modificarse la imagen original para alterar la decisión del modelo. Estas explicaciones no solo permiten identificar las regiones responsables de la clasificación actual, sino también aquellas que, al modificarse,

inducen un cambio en la predicción hacia la clase contrafactual deseada.

En contraste con los enfoques tradicionales de explicabilidad, como los métodos basados en perturbaciones locales o los ejemplos adversarios, los métodos contrafactuales no introducen ruido o cambios imperceptibles, sino que generan imágenes coherentes desde el punto de vista semántico y perceptual. Mientras que una perturbación adversaria puede cambiar la predicción del modelo mediante una alteración visual mínima pero sin sentido para un humano, los ejemplos contrafactuales visuales producen modificaciones que son comprensibles e interpretables, facilitando una mejor comprensión del comportamiento del modelo.

La literatura sobre ejemplos y explicaciones contrafactuales es extensa y abarca múltiples dominios [Verma et al., 2021, Guidotti, 2022], pero en esta sección se hace enfoque exclusivamente en aquellos métodos diseñados para tareas de clasificación de imágenes, que generan explicaciones visuales claras y plausibles desde el punto de vista semántico.

Generación de Mapas de Atribución Contrafactuales: Estos métodos generan explicaciones visuales que destacan regiones de la imagen de entrada que son informativas para una clase objetivo c (por ejemplo, la clase predicha) y, simultáneamente, no lo son para una clase contrafactual c_{counter} elegida por el usuario.

En [Goyal et al., 2019], se propone un enfoque que, dada una colección de imágenes en el momento de prueba y una clase contrafactual c_{counter}, busca reemplazar las regiones mínimas posibles de una imagen contrafactual x_{counter} con regiones de la imagen original x (clasificada como c). El objetivo es construir una nueva imagen x_c tal que el modelo f clasifique x_c como perteneciente a la clase c_{counter}.

Para reducir la complejidad del espacio de búsqueda de estas combinaciones entre x y x_{counter}, se utilizan mapas de características $F(x)$ y $F(x_{\text{counter}})$ extraídos a una resolución más baja (por ejemplo, 16×16 en lugar de 256×256 en el espacio de píxeles).

La predicción del modelo f sobre una entrada x se descompone como $f(x) = F_{\text{top}}(F(x))$, donde F extrae las características y F_{top} realiza la predicción. La salida del modelo para una clase específica c se representa como $f_c(x) = F^c_{\text{top}}(F(x))$.

Para construir x_c, se introduce una matriz de permutación P, que permite reorganizar y alinear espacialmente las regiones de $F(x_{\text{counter}})$ con las de $F(x)$. Además, se define una máscara binaria a que determina qué regiones espaciales de $F(x)$ deben preservarse y cuáles deben reemplazarse por regiones provenientes de $F(x_{\text{counter}})$. El mapa de características final optimizado se define como:

$$F(x_c) = (1 - a)F(x) + aPF(x_{\text{counter}})$$

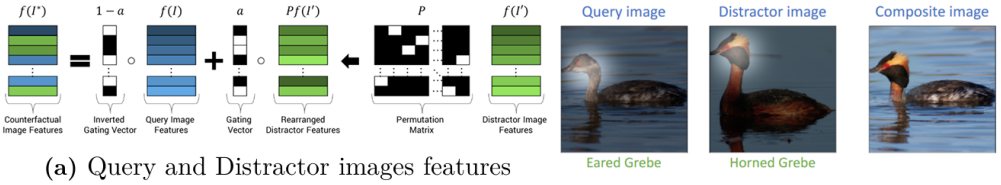
La Figura 2.3 ilustra esta operación de combinación de características.

Luego, proponen una búsqueda exhaustiva (ver Figura 2.3c) sobre todas las permutaciones de P, restringiendo la máscara binaria a a ser one-hot, o una versión relajada reparametrizando a y P usando un softmax, lo que permite la optimización basada en gradientes. La máscara a se remuestrea al tamaño de entrada (similar a [Zhou et al., 2016]) para producir el mapa de explicación final o generar un cuadro delimitador. Aunque esto captura información relevante para el clasificador dentro de la distribución de imágenes reales, deriva regiones de interés de manera heurística y aún requiere una imagen contrafactual para la comparación. Basándose en este trabajo, en [Wang and Vasconcelos, 2020] los autores generalizan la generación de explicaciones visuales contrafactuales sin requerir una colección de imágenes en la fase de testing. De manera similar, utilizan mapas de características de capas superiores (a baja resolución) pero se centran en características que son informativas para la clase de entrada c y no informativas para una clase contrafactual elegida c_{counter}. Su método propuesto es similar a GradCAM [Selvaraju et al., 2017], ya que retropropagan gradientes desde la salida del modelo hasta los mapas de características especificados. El mapa de atribución a la escala del mapa de características para x, discriminando la clase predicha c contra la clase contrafactual c_{counter}, se define como:

$$E_F(x, c, c_{\text{counter}}) = a_F(f_c(x)) \cdot \overline{a}_F(f_{c_{\text{counter}}}(x)) \cdot a_F(s(x))$$

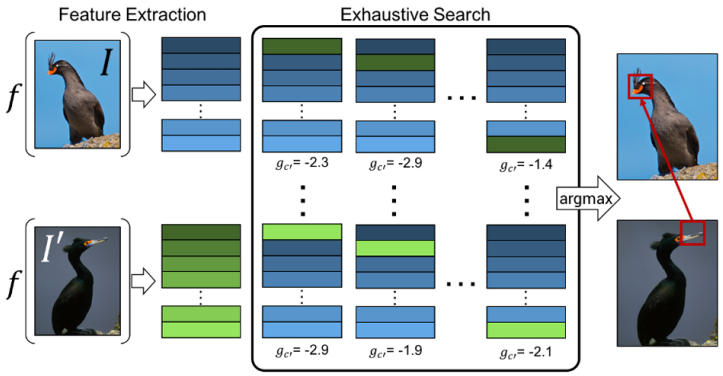
donde a_F es el producto escalar entre las derivadas parciales de la salida del modelo (para una clase específica c o c_{counter}) con respecto al mapa de características $F(x)$ (es decir,

$\partial f_c(x)/\partial F(x)$) y las activaciones del mapa de características; $\overline{a}_F$ es el complemento de a_F; y s asigna una puntuación de confianza ($s(x) \in [0, 1]$). La explicación visual en el tamaño de entrada se calcula mediante una técnica de segmentación utilizando un umbral T.



(a) Query and Distractor images features combination

(b) Counterfactual explanation



(c) Exhaustive best-edit search.

Figura 2.3: Explicaciones contrafactuales visuales utilizando imágenes distractoras [Goyal et al., 2019]. (a) Perturbación del mapa de características de la imagen input (query) usando el mapa de características de la imagen distractora. (b) De izquierda a derecha: la imagen input con la región más discriminativa iluminada, la imagen distractora también con su región más importante iluminada; y la imagen compuesta generada combinando la cabeza del animal de la imagen distractora con el cuerpo del animal de la imagen input. (c) Revisión de todos los pares de imagenes input-distractora en diferentes zonas espaciales. La selección del par es la que maximice la probabilidad de cambiar de clase.

Optimización de Máscara con Perturbación Contrafactual: Basándose en [Fong and Vedaldi, 2017], [Dabkowski and Gal, 2017] y [Ribeiro et al., 2016], estos méto-

dos [Chang et al., 2019, Agarwal and Nguyen, 2020] tienen como objetivo resolver problemas de optimización similares, pero proponen generar perturbaciones realistas utilizando modelos generativos asociados al dominio de entrada, por ejemplo, perturbando una imagen patológica con tejido sano. En [Chang et al., 2019] se resuelve iterativamente un problema similar al de [Fong and Vedaldi, 2017] rellenando la región de perturbación con el contexto de fondo de la imagen utilizando una GAN de atención contextual [Yu et al., 2018].

En la clasificación binaria, utilizando técnicas de traducción de dominio, [Samangouei et al., 2018] entrena un modelo generativo para producir, para todas las entradas, una imagen reconstruida, una imagen traducida y una región de máscara binaria. Se espera que el clasificador binario asigne una clase diferente a la imagen traducida. Sin embargo, en su marco, utilizando la arquitectura DCGAN [Radford et al., 2016], las imágenes de entrada se codifican en vectores que pasan a través de un modelo generativo para producir imágenes. Para calcular la imagen contrafactual final, utilizan una máscara binaria para combinar la entrada y la imagen traducida (para reducir los errores residuales de su proceso de generación). No obstante, todos estos enfoques aún requieren regularizaciones fuertes y heurísticas en las regiones de perturbación (es decir, la máscara) para producir mapas de explicación suaves y aceptables.

Generación Iterativa de Contrafactuales Para cada imagen de entrada, estos métodos generan de manera iterativa una imagen contrafactual que el clasificador asigna a una clase distinta, asegurando además que dicha imagen pertenezca a la distribución de datos de esa nueva clase. En [Dhurandhar et al., 2018], se optimizan dos perturbaciones δ para resaltar, respectivamente, las regiones mínimas de la entrada que son suficientes para mantener la misma clasificación (es decir, pertinentes positivos) y las regiones mínimas que, si se añaden, harían que la clasificación cambie (es decir, pertinentes negativos). Para generar perturbaciones realistas dentro de la variedad de imágenes reales, primero entrenan un autoencoder convolucional [Mousavi et al., 2017] y luego lo utilizan para restringir la perturbación durante la optimización. Otra línea de trabajo explica las decisiones de un modelo de clasificación entrenado f generando directamente ejemplos a lo largo de la distribución de datos. Utilizando un codificador preentrenado e que transforma las

imágenes de entrada x a una representación latente z, junto con un modelo generativo preentrenado g (por ejemplo, una GAN) que calcula la transformación inversa, xGEM [Joshi et al., 2018] optimiza el vector latente z para producir imágenes contrafactuales que pertenecen a la distribución aproximada de imágenes reales (mediante modelado generativo entrenado). Su objetivo es encontrar z que minimice:

$$z^* = \arg\min_z L_d(x, g(z)) + \lambda L_f(f(g(z)), t)$$

donde L_d impone la proximidad entre la entrada x y la imagen generada $g(z)$, y L_f asegura que la imagen generada sea clasificada en una clase objetivo dada t por el modelo f. El ejemplo contrafactual es entonces $x_c = g(z^*)$.

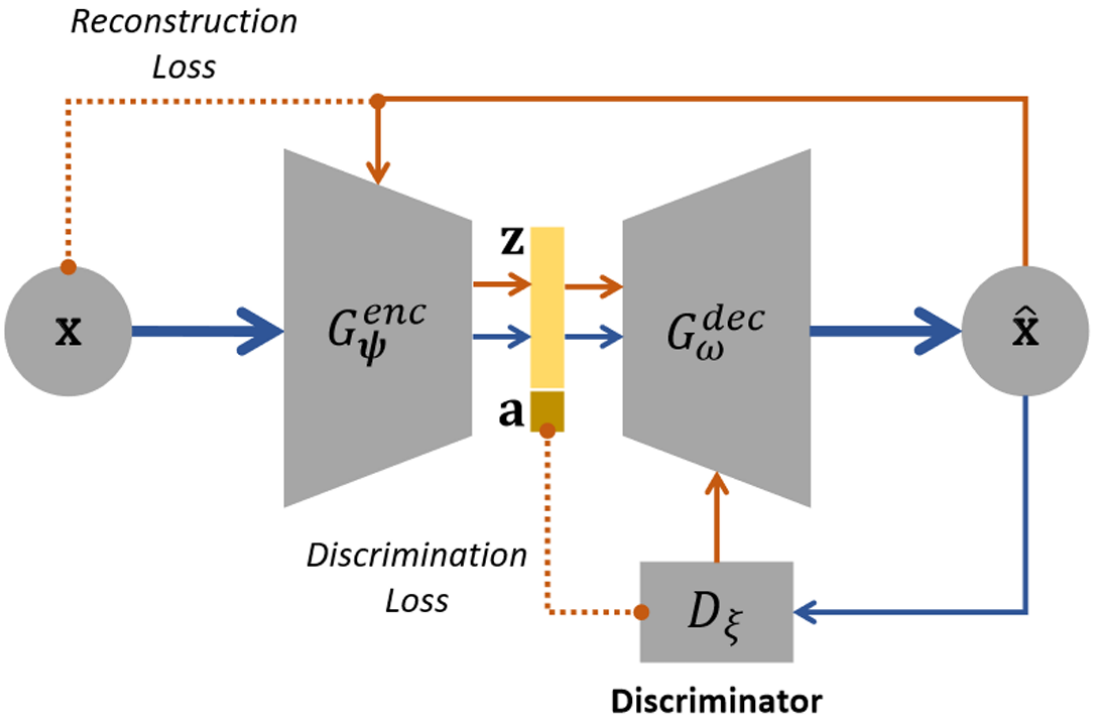


Figura 2.4: Optimización de un espacio latente informado por atributos en un modelo generativo [Yang et al., 2021].

De manera similar, en [Yang et al., 2021, Liu et al., 2019b] se buscan vectores latentes óptimos

compuestos por características brutas (no interpretables) z y características de atributos disponibles a (por ejemplo, color de cabello, género, bigote para atributos faciales). Como se ilustra en la Figura 2.4, la manipulación de características de atributos requiere modelos generativos específicos.

Más recientemente, DiVE [Rodriguez et al., 2021] optimiza múltiples perturbaciones dispersas en el espacio latente de un modelo generativo entrenado para producir contrafactuales diversos. Para fomentar la diversidad en los ejemplos generados, utilizan un β-TCVAE [Chen et al., 2019], conocido por producir representaciones latentes desentrelazadas (disentangled latent representations), para generar contrafactuales diversos. Además, imponen que diferentes contrafactuales dependan de atributos no triviales con respecto a la clase que se está explicando. Estos métodos proporcionan información sobre cómo modificar la entrada para cambiar la clasificación, ya sea para producir un prototipo de la clase predicha (aumentando la puntuación de confianza del modelo) o para generar ejemplos contrafactuales. Sin embargo, no son adecuados para calcular mapas de atribución. De hecho, al manipular vectores en el espacio latente (con pérdida de información espacial y detalles finos), los ejemplos generados a menudo difieren significativamente de la entrada, exhibiendo errores de reconstrucción (por ejemplo, imágenes borrosas, fondos o texturas diferentes, detalles faltantes). Además, los métodos optimizados en atributos conocidos permiten estudiar el impacto de cada atributo en el clasificador por separado. Sin embargo, estos atributos requieren anotaciones adicionales, que no siempre están disponibles, particularmente en el análisis de imágenes médicas.

A continuación en la Tabla 2.1 se presenta una comparación global de los principales métodos contrafactuales reportados en la literatura. La tabla resume sus características en cuanto al tipo de explicación que ofrecen (local o global) y si poseen un enfoque agnóstico al modelo, lo que permite distinguir entre técnicas dependientes de una arquitectura específica y aquellas que pueden aplicarse de manera más general.

Tabla 2.1: Comparación global de métodos contrafactuales.

Método	**Tipo Expl.**	**Model Agnostic**
Counterfactual Example [Goyal et al., 2019]	Local	✗
Counterfactual GradCAM [Wang and Vasconcelos, 2020]	Local	✗
CA-FIDO [Chang et al., 2019]	Local	✓
Pertinent++ [Dhurandhar et al., 2018]	Local	✓
xGEMs [Joshi et al., 2018]	Local/Global	✓
Attribute-Informed [Yang et al., 2021]	Local/Global	✓

2.4. Explicaciones Visuales Mediante Traducción de Dominio

2.4.1. Traducción de Dominio Mediante GANs

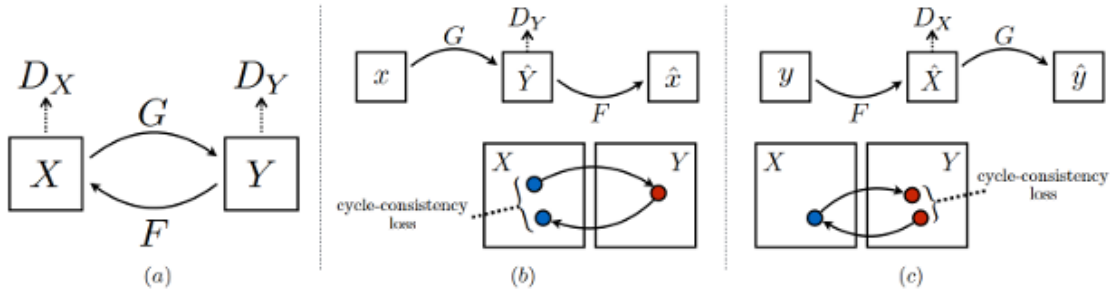
Traducción de dominio (también conocida como traducción de imagen a imagen) es una aplicación común y exitosa de las **Redes Generativas Adversarias (GANs)** [Goodfellow et al., 2014b]. Consiste en aprender una transformación (mapeo) entre dos dominios de imágenes diferentes [Isola et al., 2016].

Traducción de dominio entrelazada: Cuando se dispone de imágenes emparejadas, **Pix2Pix** [Isola et al., 2016] aprende a transformar imágenes de entrada de un dominio fuente a imágenes emparejadas en el dominio objetivo. A diferencia de las GANs tradicionales, que generan imágenes a partir de variables latentes de baja dimensión (ej: distribución normal con dimensiones entre 10 y 1000), los frameworks de traducción de dominio usan modelos generativos tipo *autoencoder* o *UNet* [Ronneberger et al., 2015] para producir una imagen a partir de una imagen de entrada. Para evitar la restricción de datos emparejados (que suelen ser escasos), **CycleGAN** [Zhu et al., 2020] introduce una restricción de consistencia cíclica que fuerza una proximidad entre dominios (dos dominios en su caso), promoviendo consistencia en la traducción y retro-traducción. Para aprender una transformación uno a uno entre dos dominios (ej: fotos ↔ pin-

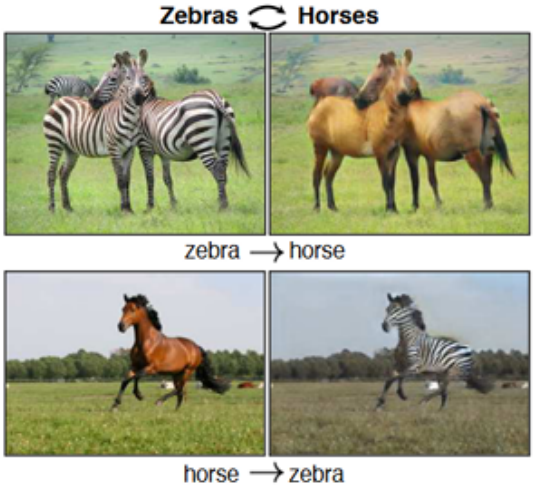
turas, cambios de estaciones, caballo ↔ cebra, imágenes sanas ↔ patológicas, transferencia de tipos de imágenes médicas), usan dos pares de generador-discriminador: (g_1, D_1) y (g_2, D_2). Dada una entrada x_1 del dominio 1 (o x_2 del dominio 2), g_1 traslada x_1 al dominio 2 (y g_2 traslada x_2 al dominio 1), intentando engañar al discriminador específico D_2 (o D_1). Para garantizar consistencia al aplicar la transformación inversa (vía g_2 o g_1), minimizan $||x_1 - g_2(g_1(x_1))||_1$ (el *término de consistencia cíclica*). La Figura 2.5a muestra las transformaciones, y la 2.5b ejemplos de traducción entre caballos y cebras.

Trabajos recientes [Lample et al., 2018, Choi et al., 2018, Yu et al., 2020, Xiao et al., 2018, Yu et al., 2019, He et al., 2018, Liu et al., 2018] extienden esta transformación a traducción multi-dominio (ej: fotos a pinturas de varios artistas, transferencia de múltiples estaciones) o edición de atributos faciales (ej: añadir gafas, bigote, cambiar peinado, color de cabello, edad o sexo). En estos enfoques, un único par generador-discriminador permite producir diferentes mapeos. Comparado con CycleGAN [Zhu et al., 2020], el generador se condiciona con el dominio objetivo o atributos a nivel de imagen [Choi et al., 2018] o en el espacio latente [Xiao et al., 2018, He et al., 2018], y se añade un módulo para clasificar el dominio (ej: un modelo de clasificación [He et al., 2018, Liu et al., 2019a] o una cabecera de clasificación en el discriminador [Choi et al., 2018]; ver Figuras 2.6).

Traducción de dominio desentrelazada: A diferencia de trabajos anteriores, estas técnicas buscan separar el contenido de la imagen (ej: características generales, pose de un rostro) del **estilo** (ej: detalles finos, colores). Una revisión exhaustiva de métodos de desentrelazado se encuentra en [Liu et al., 2022]; aquí nos enfocamos en aplicaciones para traducción de dominios.

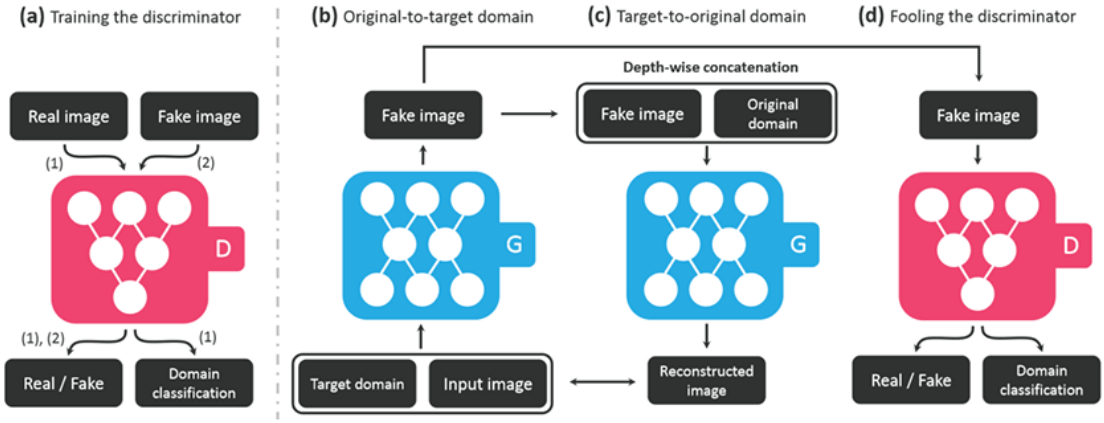


(a) Description of the CycleGAN mapping functions

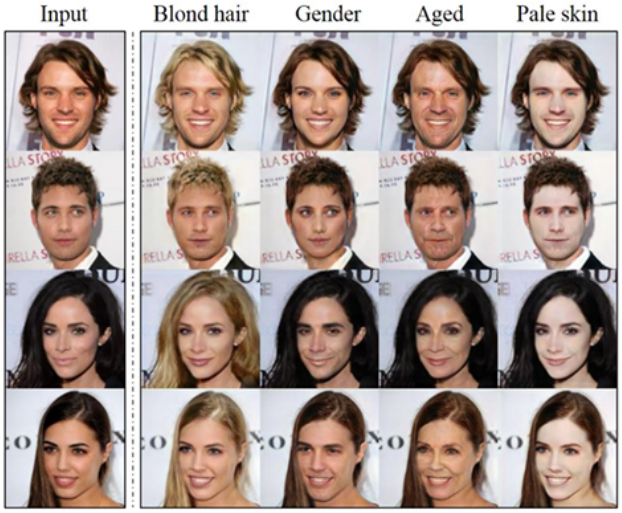


(b) Illustration of image-to-image translation

Figura 2.5: Vista general del método CycleGAN [Zhu et al., 2020]. (a) De izquierda a derecha: las transformaciones entre los dominios X e Y; donde el generador traduce imágenes del dominio X al dominio Y y trata de engañar al discriminador; La consistencia cíclica ilustrada para los inputs en X traducidas a Y por el generador y mapeados de vuelta a X. (b) Ejemplo de traducciones entre los dominios X: cebras y Y: caballos.



(a) Description of the StarGAN training



(b) Illustration of image-to-image translation

Figura 2.6: Vista general del método StarGAN [Choi et al., 2018]. (a) De izquierda a derecha: el modelo discriminador D aprende a identificar imágenes reales de imágenes sintéticas y a clasificar el dominio del input real; Dada una imagen input y un dominio objetivo, el generador G traduce la imagen en el dominio objetivo. Luego, dada esta imagen generada y el dominio original, el generador transforma de vuelta la imagen generada a la imagen original (imagen reconstruida). (b) Ejemplo de traducción de imágenes a diferentes dominios objetivos (cabello rubio, cambio de edad, cambio de género, etc).

Para cada dominio i (generalmente 2), **MUNIT** [Huang et al., 2018] y **DRIT** [Lee et al., 2018] entrenan dos *encoders*:

- e_i^c y e_i^s, que extraen el contenido c_i y el estilo s_i de una imagen de entrada x, respectivamente.
- Luego, intercambian contenido y estilo de imágenes de distintos dominios para realizar la traducción (vía un generador g_i), por ejemplo: $x_{1\to 2} = g_2(c_1, s_2)$.
- También promueven generación multimodal muestreando códigos de estilo de un prior Gaussiano ($s_i \sim N(0,1)$) y minimizando una divergencia KL junto con la distancia $||s_i - e_i^s(g_i(c_{1-i}, s_i))||_p$ (con $p \in \{1,2\}$).

En la fase de prueba, dada una entrada x del dominio $1-i$, se generan traducciones diversas en el dominio i muestreando múltiples $s_i \sim N(0,1)$ o extrayendo estilos $e_i^s(x_i)$ de una colección de imágenes del dominio i.

[Yu et al., 2019, Choi et al., 2020] amplían este enfoque incorporando un encoder de dominio. Por ejemplo, **DMIT** [Yu et al., 2019] extrae contenido, estilo y dominio de cada imagen, y los intercambia para realizar mapeos múltiples. **StarGAN v2** [Choi et al., 2020] introduce cabezales múltiples en el encoder de estilo y el discriminador (uno por dominio). Además, utiliza un mapeador de estilos multi-cabezal (similar a [Karras et al., 2019]) para aumentar la diversidad en la generación y mejorar la separación del espacio de estilos.

La Tabla 2.2 presenta una comparación entre distintos métodos de traducción y manipulación de imágenes basados en GANs. Se destacan sus capacidades en relación con el uso de datos no apareados, el soporte para dominios múltiples, la manipulación simultánea de múltiples atributos, así como aspectos relevantes como diversidad, disentanglement y los mecanismos de inyección generativa utilizados. Esta comparación permite identificar las fortalezas y limitaciones de cada enfoque, proporcionando una visión clara de las tendencias y avances más significativos en el área.

Tabla 2.2: Comparación global de métodos de traducción de dominios. Se indica si la técnica funciona con datos no emparejados; si puede traducir imágenes a múltiples dominios; si puede cambiar (o traducir) múltiples características de atributos simultáneamente; y si puede generar salidas diversas para una única entrada (y un dominio objetivo único). También destacamos si el método utiliza una estrategia de desentrelazamiento (y de qué tipo) y qué código adicional se inyecta en el modelo generativo para guiar la traducción de imágenes. «Cont.» es la abreviatura de contenido; «Est.» de estilo; «Dom.» de dominio; y «Div.» de diversidad.

Method	**Unpaired data**	**Multidomain**	**Multiattributes (Simult.)**	**Diversity**	**Disentangled**	**Gen. Injection**
Pix2Pix [Isola et al., 2016]	–	–	–	–	–	–
CycleGAN [Zhu et al., 2020]	✓	–	–	–	–	–
UNIT [Liu et al., 2018]	✓	✓	–	–	–	–
ELEGANT [Xiao et al., 2018]	✓	–	✓	–	–	–
StarGAN [Choi et al., 2018]	✓	✓	✓	–	–	–
FaderNet [Lample et al., 2018]	–	–	✓	–	–	–
AttGAN [He et al., 2018]	✓	–	✓	–	–	Dom.
STGAN [Liu et al., 2019b]	✓	–	✓	–	–	Dom.
SingleGAN [Yu et al., 2018] (1)	✓	–	–	✓	–	–
SingleGAN [Yu et al., 2018] (2)	✓	–	–	–	–	Dom. + Div.
SMIT [Romero et al., 2019]	✓	–	–	–	–	Dom. + Div.
SDIT [Wang and Vasconcelos, 2020]	✓	–	–	–	–	Dom. + Div.
MUNIT [Huang et al., 2018]	✓	–	–	✓	Cont./Sty.	Div.
DRIT [Lee et al., 2018]	✓	–	–	✓	Cont./Sty.	Div.
DMIT [Yu et al., 2019]	✓	–	–	✓	Cont./Sty./Dom.	Div.
StarGAN v2 [Choi et al., 2020]	✓	✓	–	✓	Cont./Sty./Dom.	Div.

2.4.2. Manipulación de Atributos de Imágenes

Otra línea de investigación se centra en manipular atributos de imágenes mediante el estudio del **espacio latente de modelos generativos**. El objetivo es cambiar atributos (por ejemplo, para rostros humanos: color de cabello, gafas, edad) de manera separada, continua o discreta.

No supervisado:

Al explorar el espacio latente de GANs entrenadas, los autores de [Radford et al., 2016] muestran que una interpolación lineal entre dos puntos de este espacio, o seguir una dirección específica asociada a un atributo, produce transformaciones suaves y realistas en las imágenes generadas.

Los primeros enfoques en este ámbito se centran en desentrelazar el espacio latente de modelos generativos, buscando representaciones factorizadas e interpretables. Por ejemplo, en [Higgins et al., 2016] se ajusta la importancia del término KL (Kullback-Leibler) en un VAE para fomentar la independencia entre los factores latentes.

Por otro lado, en [Chen et al., 2016] se introduce una red codificadora adicional Q, que se entrena para maximizar la información mutua entre un subconjunto c del código latente $[z, c]$ y la codificación de la imagen generada $Q(g(z, c))$, donde g es el modelo generativo.

En [Härkönen et al., 2020], se propone una técnica para manipular el espacio latente mediante el uso de **PCA** en la primera capa del generador. La base obtenida se transfiere al espacio latente mediante regresión lineal, lo que permite editar imágenes moviéndose a lo largo de esta base latente.

Una estrategia distinta es presentada en [Esser et al., 2020], donde se entrena un modelo de traducción invertible. Este modelo busca desentrelazar el espacio latente de un clasificador o autoencoder en conceptos semánticos interpretables, demostrando que las modificaciones en dicho espacio inducen cambios coherentes en las imágenes generadas.

Finalmente, en [Voynov and Babenko, 2020] se descubren direcciones semánticamente significativas en el espacio latente de una GAN entrenada. Para ello, se entrena una matriz de pesos

que permite aplicar desplazamientos de distintas magnitudes a lo largo de una dirección. Adicionalmente, se entrena un modelo reconstructor que, dada una imagen original y su versión desplazada, predice la dirección y la magnitud del cambio aplicado.

En contraste, los frameworks **StyleGAN** [Karras et al., 2019, Karras et al., 2020a, Karras et al., 2020b] introducen una arquitectura generadora alternativa para sintetizar imágenes. Su generador comienza con una entrada constante aprendida (en lugar de un ruido muestreado en $N(0,1)$), y luego los vectores de estilo, generados a partir del código latente, se pasan y actualizan en cada capa de convolución. Así, el generador puede controlar características de diferentes escalas al generar la imagen. Para producir el código latente, mapean el código de entrada muestreado $z \sim N(0,1)$ (comúnmente usado en GANs) a otro espacio latente W. El código latente w se proporciona a múltiples módulos MLP para producir los vectores de estilo para cada escala. Esta estrategia, combinada con la inyección de ruido, permite separar atributos de alto nivel (ej: formas gruesas, pose) de detalles finos. La arquitectura StyleGAN se muestra en la Figura 2.7.

Una serie de trabajos [Karras et al., 2019, Karras et al., 2020b, Collins et al., 2020, Abdal et al., 2020, Chong et al., 2021, Wu et al., 2020, Tov et al., 2021] estudian las propiedades de desentrelazado del espacio latente W o del espacio de estilo S de StyleGAN. Para manipular atributos de imágenes o mezclar estilos entre imágenes, diferentes técnicas proyectan imágenes reales en el espacio latente W (o una versión continua y ampliada W^+), entrenando un codificador específico [Pidhorskyi et al., 2020, Tov et al., 2021] o resolviendo un problema de optimización en W^+ [Abdal et al., 2020]. Basados en StyleGAN [Karras et al., 2019], otros trabajos [Zhu et al., 2020, Liu et al., 2022] modifican ligeramente la arquitectura para reforzar el desentrelazado.

Supervisado:

Otra línea de trabajo manipula atributos de imágenes utilizando cierto grado de supervisión. Los frameworks de traducción de dominio utilizados para traducción multi-dominio [Xiao et al., 2018, Choi et al., 2018, He et al., 2018, Liu et al., 2019a] pueden producir cambios que solo afectan ciertos atributos aprendidos durante el entrenamiento. Sin embargo,

solo producen traducciones hacia atributos opuestos (ej: el rostro de una persona joven o mayor) y no pueden generar un estado intermedio entre las dos clases. En contraste, [Goetschalckx et al., 2019, Jahanian et al., 2020, Plumerault et al., 2020, Yao et al., 2021] usan supervisión para encontrar direcciones interpretables en el espacio latente de GANs. Para cada transformación de imagen, optimizan un modelo lineal o no lineal para encontrar una ruta en el espacio latente tal que moverse a lo largo de esta ruta induzca un cambio continuo en la imagen.

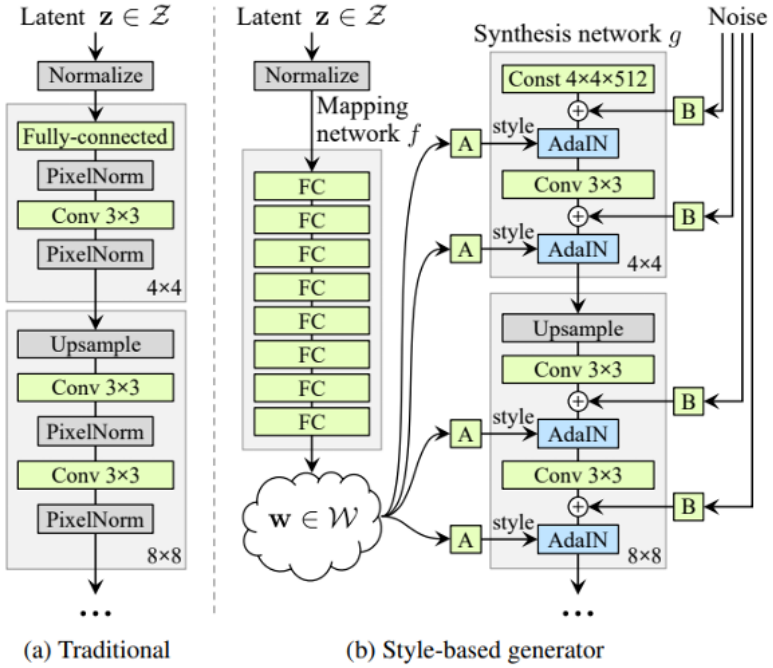


Figura 2.7: Comparación de una GAN común y una arquitectura StyleGAN [Karras et al., 2019]. (a) Arquitectura de GAN tradicional que genera imágenes directamente desde un vector latente a través de sucesivos upsamplings y operaciones de convolución. (b) La arquitectura de generador basada en estilos (style-based), donde el código latente es primero transformado a un espacio W y luego incorporado a la arquitectura del generador en diferentes niveles a través de operaciones de normalización (AdaIN).

2.4.3. Explicaciones Visuales Mediante Generaciones Contrafactuales

En el contexto de la explicación visual, estos enfoques tienen como objetivo producir un ejemplo contrafactual dentro de la distribución. La imagen generada debe parecerse a la imagen de entrada mientras muestra patrones de una clase (o dominio) diferente. En comparación con los mapas de atribución que resaltan regiones relevantes, las imágenes contrafactuales proporcionan patrones o estructuras relevantes de cada clase y muestran cómo debería cambiarse la entrada para pertenecer a otra clase.

Métodos generativos de traducción de dominio: Utilizando técnicas de traducción de dominio, estos métodos primero entrenan un modelo generativo para producir imágenes contrafactuales. En la fase de prueba, el modelo generativo produce para cada entrada una imagen contrafactual que se clasifica en una clase diferente a la de la entrada. Esta imagen contrafactual pertenece a la distribución de imágenes "reales" de la otra clase. Las diferencias entre la entrada y la imagen contrafactual generada brindan información sobre los patrones relevantes de cada clase.

Métodos de manipulación del espacio latente de GANs: En contraste con enfoques anteriores, en [Goetschalckx et al., 2019, Jahanian et al., 2020] se generan imágenes contrafactuales aprendiendo direcciones en el espacio latente de una GAN. Estas direcciones se obtienen mediante técnicas como regresión logística, máquinas de vectores de soporte o modelos de transformación más complejos. Su objetivo es modificar características semánticas relevantes de la imagen de entrada en función de una clase objetivo o una función de puntuación dada.

Otros trabajos, como [Shen et al., 2020, Yao et al., 2021], utilizan el espacio latente desentrelazado o incluso el espacio de estilo de StyleGAN [Wu et al., 2020, Karras et al., 2019]. Estas aproximaciones buscan asegurar que las generaciones alteren únicamente aquellas características relevantes para la clasificación de atributos específicos. Por ejemplo, en el caso de atributos faciales: color de cabello, presencia de bigote, género (hombre vs. mujer) o uso de gafas. En estos escenarios, todos los atributos están disponibles para entrenar el modelo de transformación latente.

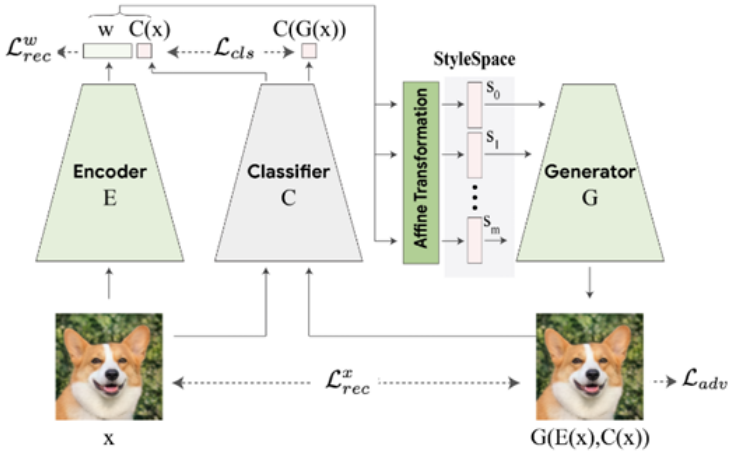
Sin embargo, cuando se enfrentan tareas de clasificación que implican múltiples atributos simultáneamente, por ejemplo, clasificar entre personas jóvenes y mayores, estos métodos tienden a modificar todos los atributos relevantes al mismo tiempo. Esto también ocurre en técnicas de traducción de dominio o en la generación iterativa de contrafactuales, donde los ejemplos generados progresivamente difieren en varios aspectos de la imagen original.

Más recientemente, **StyleEx** [Lang et al., 2021] propone una estrategia alternativa. Entrenan un modelo StyleGAN condicionado por la salida de un clasificador binario, con el fin de forzar al espacio de estilo desentrelazado a capturar mejor los atributos relevantes para la clasificación (ver el marco de optimización en la Figura 2.8a).

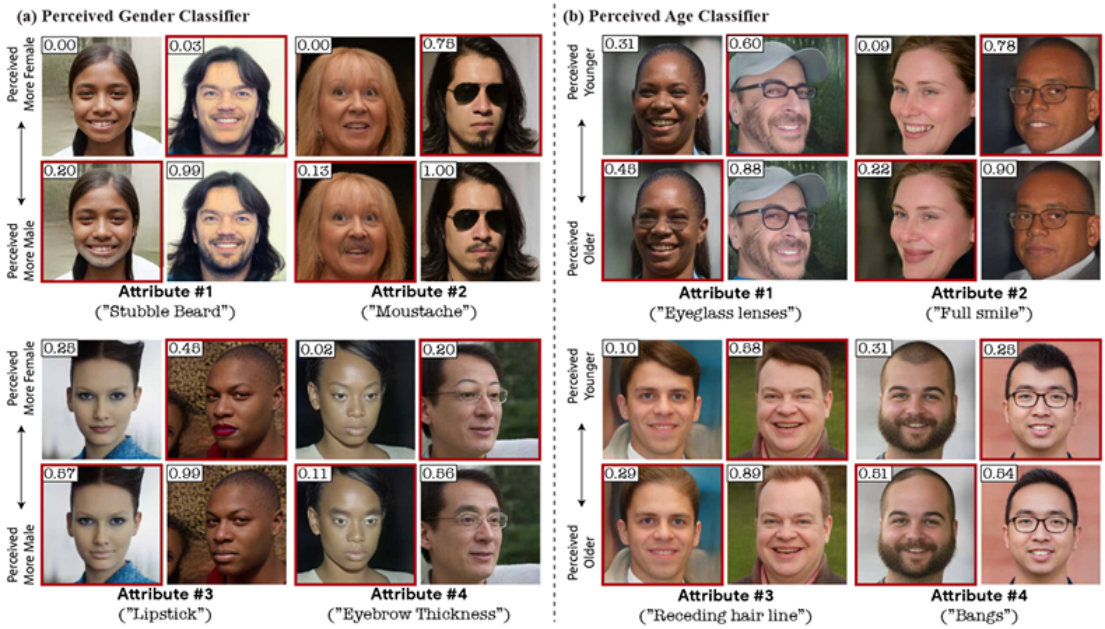
Posteriormente, identifican las dimensiones del espacio de estilo que afectan significativamente la puntuación del clasificador. Demuestran que estos vectores de estilo seleccionados corresponden a atributos desentrelazados y relevantes para la clasificación (ver Figura 2.8b).

Finalmente, utilizan un modelo codificador para proyectar imágenes reales en el espacio de estilo. De este modo, pueden modificar un atributo a la vez y evaluar su impacto directo en la predicción del clasificador.

Aunque el framework StyleGAN produce imágenes de alta calidad en comparación con técnicas anteriores de GAN, y a pesar de los esfuerzos para proyectar imágenes reales en el espacio latente de StyleGAN, algunos contenidos de la imagen (por ejemplo, elementos en el fondo, texturas de objetos o detalles) difieren. **La clasificación latente no traduce perfectamente lo que el modelo de clasificación de imágenes ha aprendido. Esto induce un sesgo en la explicación**. Estos métodos están más adaptados para proporcionar información sobre los atributos relevantes de cada clase en lugar de producir una explicación post-hoc de la decisión de un clasificador (sobre una entrada específica).



(a) Training framework



(b) Manipulation of the top detected attributes.

Figura 2.8: Arquitectura general de Stylex [Lang et al., 2021]. (a) Entrenamiento: El encoder E codifica el input x en espacio de estilos de styleGAN. El vector codificado w es concatenado con la predicción del input x. El vector resultante aprende transformaciones afines para producir vectores de estilo para cada escala de la ruta generativa. El generador G trata de reconstruir el input y preservar la decisión de clasificación. (b) Para el género y la edad, las transformaciones a través de diferentes direcciones de atributos detectados son visualmente realistas y relevantes para la tarea de clasificación.

En la Tabla 2.3 presenta una comparación global de los métodos contrafactuales de traducción de dominio. En ella se describen las características principales de cada técnica, incluyendo el tipo de explicación generada (local, global o únicamente informativa), si son agnósticas al modelo o a los datos, y el tipo de estrategia empleada para la generación de contrafactuales. Esta síntesis permite identificar qué métodos ofrecen explicaciones aplicables directamente sobre clasificadores o regresores, cuáles requieren acceso a estructuras internas del modelo y qué aproximaciones dependen exclusivamente de bases de datos o de la arquitectura del generador.

Tabla 2.3: Comparación global de métodos de contrafactuales de traducción de dominio. Para cada método, indicamos si la explicación visual es local o global (Nota: «Insights only» significa que la técnica no se aplica a un clasificador, sino que proporciona información sobre la tarea); si el usuario tiene acceso a las estructuras internas del modelo (y cuáles) o no; si se requiere datos adicionales o solo los datos probados; y qué tipo de técnica se utiliza. «NN as Encoder» implica que la red neuronal estudiada se usa como parte codificadora del modelo generativo. «Gen.» es la abreviatura de modelo generador.

Method	**Expl. Type**	**Model Agnostic**	**Data Agnostic**	**Technique**
Insights via CycleGAN [Narayanaswamy et al., 2020]	Insights only	✗	Database (Gen. training)	Not Explaining a Classifier
VA-GAN [Baumgartner et al., 2018]	Insights only	✗	Database (Gen. training)	Not Explaining a Classifier
VR-GAN [Bigolin Lanfredi et al., 2019]	Insights only	✓	Database (Gen. training)	Not Explaining a Regressor
BIN [Oh et al., 2021]	Local/Global	NN as Encoder	—	Additive Counterfactual Mask Generator Optim.
PE [Singla et al., 2020]	Local/Global	✓	Database (Gen. training)	Counterfactual Generator Optim.
Latent Classif. in StyleGAN [Shen et al., 2020]	Insights	✓	Database (StyleGAN training)	Latent Manipulation (not really an explanation)
StyleEx [Lang et al., 2021]	Insights	✓	Database (Enc., StyleGAN training)	Latent Manipulation

2.5. Métricas de Evaluación de Calidad de Imagen y Explicaciones Visuales

Para evaluar cuantitativamente la calidad y fidelidad de las imágenes histopatológicas generadas por el método propuesto, se emplearon diversas métricas estándar en la evaluación de modelos generativos. Entre ellas se encuentran; FID, PSNR, SSIM y LPIPS. Luego, para evaluar cualitativa y cuantitativamente las explicaciones visuales, se detalla un método clásico basado en

encuestas con personas reales, el Test de Turing Visual.

Métrica: Fréchet Inception Distance (FID)

La métrica FID [Heusel et al., 2017] se ha consolidado como una de las principales herramientas para evaluar la calidad de las imágenes generadas por arquitecturas GAN. Su objetivo es medir la similitud entre la distribución de un conjunto de imágenes generadas y la distribución de un conjunto de imágenes reales, cuantificando así el realismo y la fidelidad de las imágenes sintéticas.

En lugar de comparar imágenes individuales píxel a píxel, el FID opera comparando las distribuciones estadísticas de características extraídas de ambas colecciones de imágenes (reales y generadas) mediante una red neuronal profunda. Específicamente, se ajustan dos distribuciones Gaussianas multidimensionales a los vectores de activaciones:

- $\mathcal{N}(\mu_r, \Sigma_r)$ para las imágenes reales.
- $\mathcal{N}(\mu_g, \Sigma_g)$ para las imágenes generadas.

La métrica se define como la distancia de Wasserstein [Arjovsky et al., 2017] entre estas dos Gaussianas:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\big(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\big). \tag{2.1}$$

Donde:

- μ_r y μ_g son los vectores de medias de las características de las imágenes reales y generadas, respectivamente.
- Σ_r y Σ_g son las matrices de covarianza empíricas de dichas características.
- $\|\cdot\|_2$ denota la norma Euclidiana.
- $\text{Tr}(\cdot)$ representa la traza de una matriz.

Para extraer los vectores de características (μ y Σ), tanto las imágenes reales como las generadas se procesan a través de una red InceptionV3 pre-entrenada en el conjunto de datos

ImageNet[Deng et al., 2009].

Un valor más bajo de FID indica una mayor similitud entre las distribuciones de las imágenes reales y generadas, interpretándose como una mayor calidad y realismo de las muestras sintéticas.

Aunque el FID estándar es muy útil para comparaciones inter-modelo, presenta limitaciones cuando se aplica a dominios especializados como la histopatología. Al usar una red entrenada en imágenes naturales (ImageNet), el FID mide principalmente el realismo visual general bajo la óptica de objetos cotidianos, sin garantizar una fiel representación de las características microscópicas relevantes para el diagnóstico histopatológico. Por lo tanto, en contextos de contrafactuales o de transferencia de dominio, podría no reflejar adecuadamente la fidelidad de los detalles morfológicos específicos.

Métrica: PSNR (Peak Signal-to-Noise Ratio)

El *Peak Signal-to-Noise Ratio* (PSNR) es una métrica tradicional utilizada para medir la calidad de imágenes reconstruidas, basada en el error cuadrático medio (MSE) entre dos imágenes. Se define como:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{MAX_I^2}{\text{MSE}} \right)$$

donde MAX_I representa el valor máximo posible de un píxel (por ejemplo, 255 para imágenes de 8 bits). Aunque PSNR es ampliamente utilizado por su simplicidad y fácil interpretación, su principal limitación es que no modela adecuadamente la percepción visual humana, ya que trata todos los errores de manera uniforme, independientemente de su contexto visual [Horé and Ziou, 2010].

Métrica: SSIM (Structural Similarity Index Measure)

El *Structural Similarity Index* (SSIM), propuesto por [Wang et al., 2004], busca mejorar la correlación con la percepción humana modelando tres componentes: luminancia, contraste y es-

tructura. Su formulación es:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

donde μ_x y μ_y son las medias locales, σ_x^2 y σ_y^2 las varianzas, y σ_{xy} la covarianza entre las ventanas de imagen x e y. SSIM varía entre -1 y 1, donde 1 indica máxima similitud estructural. Esta métrica ha demostrado mayor correlación con juicios humanos que PSNR, especialmente en imágenes comprimidas y reconstruidas [Wang et al., 2004].

Métrica: LPIPS (Learned Perceptual Image Patch Similarity)

El *Learned Perceptual Image Patch Similarity* (LPIPS), introducido por [Zhang et al., 2018], es una métrica perceptual basada en redes neuronales profundas. Compara imágenes a través de distancias entre activaciones intermedias de redes convolucionales pre-entrenadas, ponderando estas diferencias con parámetros aprendidos que reflejan mejor la percepción humana. A diferencia de PSNR y SSIM, LPIPS es capaz de capturar similitudes semánticas, texturas y estilos globales. Un menor valor de LPIPS indica una mayor similitud perceptual entre las imágenes.

Aplicación en GANs

En la evaluación de modelos GAN, especialmente aquellos diseñados para tareas de reconstrucción o edición controlada de imágenes, estas métricas son ampliamente utilizadas como indicadores cuantitativos de calidad. PSNR y SSIM son útiles en tareas donde la fidelidad local es crítica, aunque pueden no reflejar adecuadamente la calidad percibida cuando se introducen variaciones estructurales sutiles pero plausibles. Por otro lado, LPIPS se ha posicionado como una métrica estándar para evaluar la calidad perceptual de imágenes generadas por GANs, ya que correlaciona fuertemente con evaluaciones humanas en tareas como superresolución, transferencia de estilo y generación de contrafactuales [Zhang et al., 2018, Isola et al., 2016]. Su uso permite cuantificar mejoras en la plausibilidad visual de muestras generadas, aspecto central en aplicaciones como medicina, arte generativo o síntesis de rostros realistas.

Test de Turing Visual

La evaluación de la calidad y el realismo de las imágenes sintéticas producidas por modelos generativos es un aspecto crucial en el desarrollo y la validación de dichas tecnologías. Más allá de las métricas cuantitativas automáticas que comparan propiedades estadísticas o estructurales de las imágenes (como FID, LPIPS, SSIM, etc.), existe la necesidad de evaluar cómo estas imágenes son percibidas por observadores humanos. En este contexto, el Test de Turing Visual [Geman et al., 2015] emerge como una metodología de evaluación fundamental.

Inspirado directamente en el célebre Test de Turing propuesto por Alan Turing en 1950 para evaluar la capacidad de una máquina de exhibir un comportamiento inteligente indistinguible del de un humano [Turing, 2009], el Test de Turing Visual adapta este paradigma al dominio de la generación de imágenes. El objetivo central del Test de Turing Visual es determinar si las imágenes generadas por un modelo computacional son perceptualmente indistinguibles de imágenes reales para un observador humano.

La implementación típica de un Test de Turing Visual consiste en presentar a un grupo de evaluadores humanos un conjunto de imágenes que contiene una mezcla aleatoria de muestras reales (extraídas del dominio de interés) y muestras sintéticas (generadas por el modelo que se desea evaluar). La tarea de los evaluadores es clasificar cada imagen presentada como “Real” o “Generada” (“Falsa”). En diseños experimentales más elaborados, se puede solicitar adicionalmente a los evaluadores que indiquen un nivel de confianza en su decisión o que proporcionen comentarios cualitativos sobre por qué consideran que una imagen es real o sintética.

Un aspecto crítico en el diseño de un Test de Turing Visual es la selección de los evaluadores. Si bien para tareas de generación de imágenes generales (por ejemplo, rostros o paisajes) podrían participar evaluadores no expertos, en dominios altamente especializados como la medicina [Chuquicusma et al., 2018] y, en particular, la histopatología, es indispensable contar con la participación de expertos del dominio (en este caso, patólogos o profesionales con experiencia en la interpretación de imágenes histológicas). La razón es que estos expertos poseen el conocimiento y la experiencia visual para detectar artefactos sutiles, inconsistencias morfológicas o

desviaciones de la plausibilidad biológica que un observador sin entrenamiento probablemente pasaría por alto. La evaluación por expertos proporciona, por lo tanto, una medida mucho más rigurosa y clínicamente relevante del realismo alcanzado por el modelo generativo.

Se considera que un modelo generativo ha tenido éxito en el Test de Turing Visual si los evaluadores humanos no pueden distinguir las imágenes sintéticas de las reales con una precisión significativamente superior al azar [Goodfellow et al., 2014b]. Es decir, si la tasa de acierto promedio de los evaluadores en la tarea de clasificación se aproxima al 50 %, indica que las imágenes generadas son perceptualmente muy similares a las reales. Por el contrario, si los evaluadores logran una precisión considerablemente mayor al 50 %, sugiere que las imágenes generadas aún contienen señales o características que permiten diferenciarlas de las auténticas.

El Test de Turing Visual no reemplaza a las métricas cuantitativas, sino que las complementa, ofreciendo una perspectiva invaluable sobre la calidad perceptual. Mientras que métricas como el FID pueden cuantificar la similitud de distribuciones en un espacio de características, el Test de Turing Visual evalúa directamente si el resultado final "engaña" al sistema visual humano, que es el árbitro final de la calidad visual en muchas aplicaciones. Determinar si las imágenes generadas son indistinguibles de las reales para los expertos del dominio es fundamental para establecer la confianza y la utilidad potencial de los modelos generativos en aplicaciones sensibles como el diagnóstico asistido, la educación médica o la investigación basada en datos sintéticos.

Capítulo 3

Metodología

Esta sección presenta la metodología desarrollada para la generación de explicaciones contrafactuales visuales en imágenes histopatológicas mediante modelos generativos. El enfoque metodológico se estructura en siete componentes principales que abordan de manera integral el problema de la explicabilidad en clasificación de imágenes médicas.

En primer lugar se estudia el trabajo relacionado en cuando a modelos generativos contrafactuales en histopatología y el dominio médico general. Se identifica la brecha existente y se le da un posicionamiento al trabajo.

En segundo lugar, se describe la selección y caracterización de los conjuntos de datos utilizados, con especial énfasis en el dataset NCT-CRC-HE-100K para imágenes de cáncer colorectal, el cual proporciona la base experimental para validar la propuesta.

Luego, se incluye la descripción de un trabajo preliminar llevado a cabo antes del desarrollo de esta tesis, en el cual se exploró la viabilidad del uso de modelos generativos, específicamente StyleGAN2-ADA, para el aumento de datos sintéticos. Este trabajo, que culminó en la publicación de un artículo científico, permitió validar la capacidad de esta arquitectura para generar imágenes histopatológicas sintéticas de alta calidad, así como identificar aspectos críticos para su futura aplicación en tareas de explicabilidad. Las lecciones extraídas de esta etapa inicial orientaron decisiones clave en el diseño de la metodología principal aquí presentada.

Posteriormente, se detalla el desarrollo del modelo generativo basado en StyleGAN2-ADA, que constituye la arquitectura central del sistema propuesto, integrando un generador condicional, un clasificador preentrenado y un encoder para la proyección de imágenes reales al espacio latente.

En quinto lugar, se aborda el desarrollo del método contrafactual, donde se presenta un algoritmo innovador que utiliza vectores de ruido fijos para mantener la coherencia estructural de las imágenes mientras explora sistemáticamente el espacio latente, generando así múltiples versiones de una misma imagen con variaciones semánticamente relevantes. Esta aproximación se complementa con técnicas de interpolación latente que permiten visualizar de forma continua las transiciones entre clases, facilitando la identificación de los cambios morfológicos específicos que influyen en la decisión del clasificador.

Luego, se aborda la exploración y análisis del espacio latente, que se enfoca en la identificación de patrones semánticos mediante técnicas de reducción de dimensionalidad y análisis de densidad. Este componente permite establecer vínculos directos entre las regiones del espacio latente y las características visuales representativas de cada clase diagnóstica.

Finalmente, se presenta el desarrollo de un software de validación con expertos clínicos, diseñado para evaluar la utilidad práctica y la interpretabilidad de las explicaciones contrafactuales generadas, proporcionando una interfaz que facilite la evaluación sistemática de los resultados por parte de profesionales médicos especializados.

3.1. Trabajo Relacionado

La patología digital, apoyada por los avances en aprendizaje profundo, ha transformado el potencial del análisis de imágenes histopatológicas para el diagnóstico y la investigación oncológica [Litjens et al., 2017, Madabhushi and Lee, 2016]. Si bien la capacidad predictiva de estos modelos es innegable, su limitada transparencia sigue siendo una barrera para la confianza y la adopción clínica [Tizhoosh and Pantanowitz, 2018]. En este contexto, la Inteligencia Artificial Explicable (XAI) busca dotar de interpretabilidad a estos sistemas. Entre los paradigmas de XAI, las explicaciones contrafactuales han emergido como una aproximación particularmente poten-

te, ya que buscan responder a la pregunta "¿Qué cambios mínimos en la imagen modificarían la predicción del modelo?"[Wachter et al., 2017, Verma et al., 2021], alineándose con el razonamiento clínico basado en escenarios hipotéticos. La síntesis de estas explicaciones de forma visualmente realista y significativa representa un desafío considerable, donde los modelos generativos profundos ofrecen una vía prometedora [Goodfellow et al., 2014b, Joshi et al., 2018].

3.1.1. Aplicaciones de Explicaciones Generativas y Contrafactuales en Imágenes Médicas

La idea de emplear modelos generativos para crear explicaciones visuales, especialmente contrafactuales, ha comenzado a materializarse en diversas áreas de la imagenología médica, sentando precedentes relevantes para el campo de la histopatología. Un ejemplo notable es el trabajo conceptualizado como “CheXplaining in Style” [Atad et al., 2022] (o enfoques análogos), que explora cómo técnicas basadas en la manipulación de estilo pueden generar explicaciones visuales para radiografías de tórax. Estos métodos buscan mostrar cómo una imagen cambiaría para reflejar la adición o eliminación de una patología específica. Para lograrlo, se apoyan en la capacidad de control fino de las arquitecturas generativas avanzadas, permitiendo así generar visualizaciones que ilustran los cambios clave que influirían en la decisión del modelo.

Más allá de las radiografías de tórax, se han explorado enfoques generativos contrafactuales en otros dominios. Por ejemplo, en dermatología, se han desarrollado métodos para generar imágenes contrafactuales de lesiones cutáneas, mostrando cómo una lesión benigna podría transformarse para ser clasificada como maligna por un modelo de IA, ayudando así a evaluar la robustez y los posibles sesgos del clasificador [Metta et al., 2024]. De manera similar, en neuroimagen, se han utilizado modelos causales profundos y técnicas generativas para simular efectos de tratamiento o visualizar cómo diferencias estructurales mínimas podrían llevar a diferentes clasificaciones diagnósticas, proporcionando inferencias contrafactuales tratables [Pawlowski et al., 2020]. Otros trabajos han empleado métodos generativos sobre imágenes de retinopatía diabética y fracturas por compresión vertebral [Atad et al., 2024] para la creación de contrafactuales, demostrando la versatilidad de los enfoques generativos para crear estas expli-

caciones visuales.

Estos métodos no solo buscan explicar decisiones específicas, sino que también abren nuevas vías para la investigación médica. Como argumentan en [Tanyel et al., 2023], la capacidad de generar escenarios contrafactuales plausibles permite a los investigadores explorar hipótesis y relaciones causales potenciales que van más allá de los datos observados. Permiten formular preguntas como "¿Cómo se vería este tejido si este marcador molecular específico estuviera ausente?" o "¿Qué cambios morfológicos se le deben hacer a un tejido benigno para clasificarse como un tejido canceroso?". Este potencial para ir **más allá de la realidad conocida** subraya la importancia de desarrollar métodos contrafactuales robustos y fiables para la investigación biomédica.

3.1.2. Estado Actual en Histopatología: Generación, Explicación y la Brecha Existente

Al dirigir la atención específicamente a la histopatología computacional, observamos un uso creciente de modelos generativos, aunque predominantemente enfocados en otras tareas distintas a la generación de explicaciones contrafactuales. Las GANs, por ejemplo, han sido ampliamente utilizadas para la síntesis de imágenes histopatológicas realistas con el fin de aumentar conjuntos de datos, lo cual es crucial dada la frecuente escasez de datos anotados en este dominio [Wei et al., 2019, Xue et al., 2021, Tellez et al., 2019, Muñoz et al., 2025a]. También se han aplicado para la normalización de tinciones o la traducción entre dominios (ver Figura 3.1), lo cual indirectamente puede ayudar a la robustez de los modelos pero no constituye una explicación contrafactual por sí misma [Sloboda et al., 2024].

En cuanto a la explicabilidad en histopatología, los métodos predominantes siguen siendo los basados en saliency maps (como CAM o Grad-CAM) [Selvaraju et al., 2017, Chattopadhay et al., 2018] que identifican regiones importantes pero sin ofrecer una perspectiva contrafactual ni generar ejemplos visuales alternativos, entre otras limitaciones que ya fueron discutidas. Si bien existen algunos trabajos que exploran distintos tipos de explicaciones (detalladas en la Sección 2), la generación de imágenes histopatológicas contrafactuales realistas,

coherentes y semánticamente controlables sigue siendo un área no desarrollada. La complejidad de las imágenes histopatológicas requiere un nivel de control en la generación que va más allá de lo ofrecido por muchos enfoques genéricos. **La inspiración proveniente de arquitecturas con capacidades avanzadas de desentrelazamiento y edición semántica (como las exploradas conceptualmente en trabajos tipo StyleX [Karras et al., 2020b, Lang et al., 2021]) resulta particularmente relevante aquí, ya que apuntan hacia el tipo de control fino sobre atributos visuales a distintas escalas que sería necesario para generar contrafactuales histopatológicos significativos**. Sin embargo, la aplicación y adaptación de tales principios para la generación contrafactual en este dominio específico es aún incipiente.

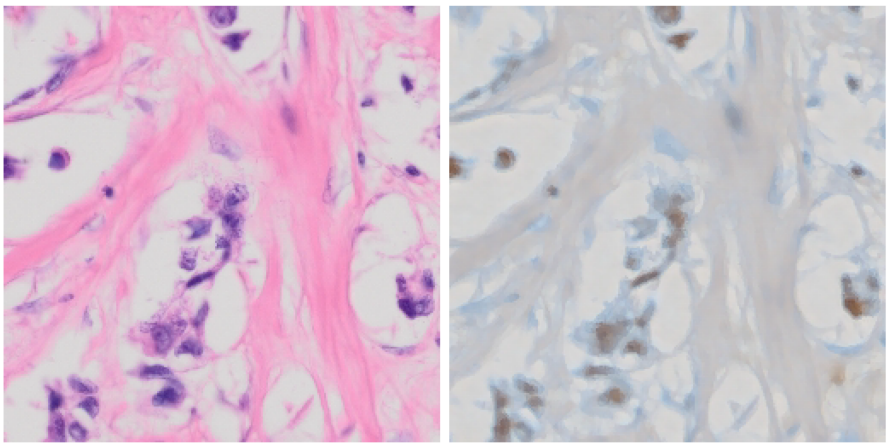


Figura 3.1: **Ejemplo de traducción de dominio.** Demostración de la conversión de dominio realizada por el modelo propuesto en [Sloboda et al., 2024] de una imagen histopatológica con tinción H&E a una imagen con tinción p63. (a) Imagen resaltada con tinción H&E sin editar que no presenta células mioepiteliales. (b) Imagen resaltada con tinción p63 tras la transformación.

A pesar de los avances en modelos generativos (particularmente aquellos con control de estilo, inspirados por arquitecturas como StyleGAN y trabajos de edición semántica como los conceptualizados en StyleX [Karras et al., 2020a, Lang et al., 2021]) y el demostrado poten-

cial de las explicaciones contrafactuales en otras áreas médicas (como lo ilustra CheXplaining [Atad et al., 2022]), **existe una notable brecha en la literatura en cuanto a la aplicación sistemática y adaptada de métodos generativos contrafactuales específicamente para imágenes histopatológicas**. La complejidad y variabilidad morfológica única de estas imágenes –su alta resolución, la sutileza de las características diagnósticas a múltiples escalas, la variabilidad de la tinción, y la necesidad de plausibilidad clínica validada por expertos– plantea desafíos que no son directamente abordados por los métodos desarrollados para otros tipos de imágenes.

Actualmente, faltan enfoques que:

- Utilicen el poder de los modelos generativos modernos con control fino de estilo para sintetizar imágenes histopatológicas contrafactuales realistas.
- Permitan una manipulación semántica dirigida a características histomorfológicas específicas relevantes para el diagnóstico (e.g., atipia nuclear, patrones glandulares).
- Generen explicaciones que no solo cambien la predicción del modelo, sino que también representen transformaciones patológicamente plausibles y útiles para la comprensión clínica.

3.1.3. Conclusión Parcial y Posicionamiento del Trabajo

El estado actual de la investigación en xAI demuestra la madurez de los modelos predictivos en histopatología y una creciente necesidad de explicabilidad contrafactual que supere los métodos actuales. Si bien los modelos generativos y las explicaciones contrafactuales han mostrado ser prometedores por separado y en otras aplicaciones médicas, su combinación sinérgica y adaptada a los desafíos específicos de la histopatología sigue siendo un área insuficientemente explorada.

Inspirándose en las capacidades de control semántico y en la utilidad explicativa demostrada en múltiples dominios y motivada por el potencial de investigación que justifican los contrafactuales en el ámbito médico [Tanyel et al., 2023], esta tesis propone un nuevo método que aborda esta área poco explorada. El objetivo es desarrollar y evaluar un nuevo método capaz de generar explicaciones contrafactuales visuales en histopatología que sean realistas, semánticamente

controlables y clínicamente relevantes, contribuyendo así a la creación de sistemas de IA más transparentes, confiables y útiles para el avance del diagnóstico y la investigación histopatológica.

Para esto se propone un método basado en tres componentes: un modelo generativo basado StyleGAN2-ADA, un encoder y un clasificador. Además, se propone un enfoque que combina: (1) una estrategia de entrenamiento en dos etapas que optimiza primero la reconstrucción de imágenes antes de introducir la tarea de clasificación, y (2) la exploración sistemática del espacio latente W para la generación de contrafactuales mediante modificaciones visuales controladas. Este enfoque aprovecha las propiedades de desentrelazamiento del espacio W y su capacidad para vincular regiones latentes con características histopatológicas específicas, proporcionando la base semántica necesaria para generar explicaciones contrafactuales clínicamente relevantes. Para validar la efectividad del método propuesto, se realizará una evaluación con expertos clínicos junto con una comparación sistemática con el enfoque "Chexplaining in Style", evaluando tanto la eficiencia computacional como la validez semántica de los contrafactuales generados.

3.2. Descripción de los Datasets

En esta sección se describen los conjuntos de datos utilizados en el desarrollo de esta investigación: **PatchCamelyon (PCam)**, **Invasive Ductal Carcinoma (IDC)**, **Breast Cancer Histopathological Annotation and Diagnosis dataset (BreCaHAD)** y **NCT-CRC-HE-100K**. El dataset PCam fue seleccionado por ser un benchmark ampliamente utilizado en tareas de clasificación de imágenes histopatológicas, destacando por su simplicidad y accesibilidad. De manera similar, el conjunto IDC ofrece facilidad de implementación gracias al reducido tamaño de sus parches, pero se diferencia por estar enfocado en una patología específica (carcinoma ductal invasivo) y en una única región anatómica (tejido mamario). En contraste, BreCaHAD aporta imágenes de resolución significativamente superior, lo que permite evaluar la capacidad del modelo para adaptarse a datos de mayor calidad visual. Finalmente, NCT-CRC-HE-100K fue incluido por su mayor relevancia clínica, su especificidad en el diagnóstico del cáncer colorrectal y

su resolución intermedia, características clave para abordar de manera más realista el problema de clasificación en contextos médicos reales.

Descripción del Conjunto de Datos PatchCamelyon (PCAM)

PatchCamelyon (PCam) es un conjunto de datos balanceado compuesto por un total de **327.680 imágenes a color** de tamaño 96×96 **píxeles**, extraídas de escaneos histopatológicos de secciones de ganglios linfáticos (ver Figura 3.2). Cada imagen está etiquetada de forma binaria para indicar la **presencia o ausencia de tejido metastásico**. Las etiquetas positivas en PCam se asignan a aquellas imágenes cuyo área central (32×32 píxeles) contiene al menos un píxel correspondiente a tejido tumoral. Este dataset fue diseñado para abordar tareas de detección de metástasis mediante clasificación binaria de imágenes y debido a su simpleza, es comparable a otros benchmarks ampliamente utilizados como CIFAR-10 o MNIST.

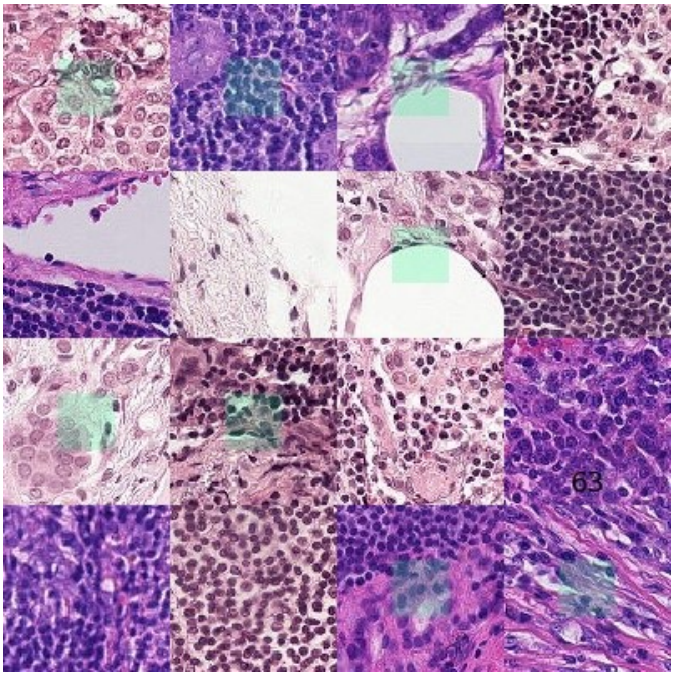


Figura 3.2: **Imágenes extraídas del dataset PatchCamelyon (PCAM).** Las imágenes cancerosas se representan con un área verde central que indica la presencia de al menos un píxel de tejido tumoral.

Estructura del Dataset

El conjunto de datos está dividido en tres subconjuntos principales, con las siguientes distribuciones:

- **Entrenamiento:** 262.144 imágenes (80 % del total).
- **Validación:** 32.768 imágenes (10 % del total).
- **Prueba:** 32.768 imágenes (10 % del total).

Es importante destacar que no existe solapamiento entre las imágenes de los diferentes subconjuntos. Cada división mantiene una distribución (50/50) entre ejemplos positivos y negativos, lo que asegura un balance de clases durante el entrenamiento y evaluación de los modelos.

Descripción del Conjunto de Datos Invasive Ductal Carcinoma (IDC)

El conjunto de datos *Invasive Ductal Carcinoma (IDC)* [Janowczyk and Madabhushi, 2016] consiste en muestras de imágenes histopatológicas derivadas de 162 imágenes de portaobjetos completos (Whole Slide Images, WSI) de biopsias de cáncer de mama, escaneadas a una magnificación de 40× utilizando un escáner digital. Este dataset fue introducido con el objetivo de apoyar el diagnóstico automatizado de subtipos de cáncer de mama, proporcionando imágenes de alta resolución anotadas para la presencia de carcinoma ductal invasivo (IDC), el subtipo más común y agresivo del cáncer de mama. El IDC representa aproximadamente el 80 % de todos los cánceres de mama invasivos y se caracteriza por la proliferación descontrolada de células epiteliales malignas que infiltran el tejido mamario circundante. Por lo tanto, la detección temprana y precisa del IDC mediante análisis histopatológico es crítica para mejorar los resultados clínicos.

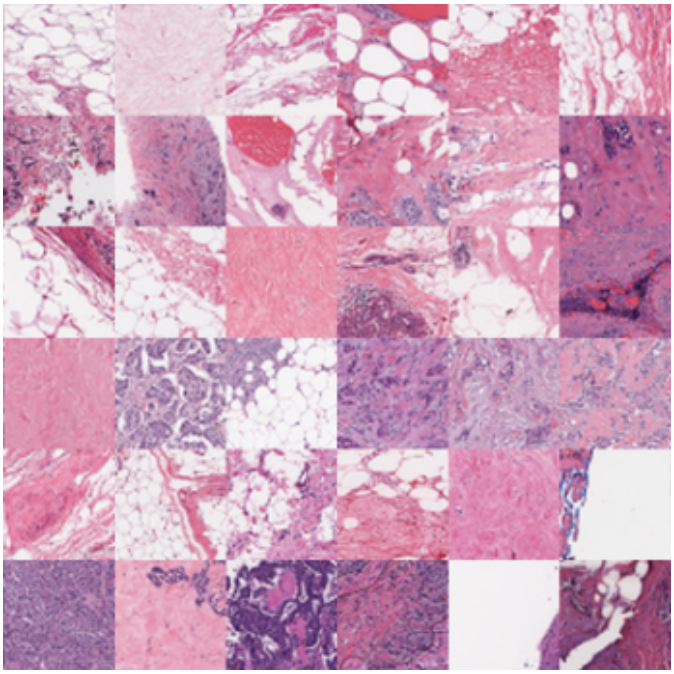


Figura 3.3: **Imágenes extraídas del dataset Invasive Ductal Carcinoma (IDC).** Las imágenes mostradas corresponden a parches reescalados a una resolución de 64 × 64.

Estructura del Dataset

En lugar de utilizar las imágenes completas, el dataset está compuesto por **277.524** parches de imagen no superpuestos, cada uno de tamaño **50×50 píxeles**, extraídos de regiones tumorales previamente anotadas. Cada parche está etiquetado de forma binaria: una etiqueta 0 indica la ausencia de IDC, mientras que una etiqueta 1 señala su presencia. De los parches extraídos, 198.738 (71,6 %) son negativos para IDC y 78.786 (28,4 %) son positivos, lo que introduce un desbalance de clases significativo, aproximadamente de 2,5:1 a favor de los casos negativos. El conjunto fue dividido aleatoriamente en subconjuntos de entrenamiento, validación y prueba en una proporción de 7:1:2, manteniendo la distribución original de clases en cada partición. El pequeño tamaño de los parches y la similitud visual inherente entre regiones benignas y malignas aumentan la dificultad de la tarea de clasificación, haciendo de este dataset un recurso particularmente adecuado para evaluar tanto el rendimiento como la robustez de modelos de aprendizaje profundo.

Relevancia de PCam y IDC para la Investigación

Los conjuntos de datos PatchCamelyon (PCam) e Invasive Ductal Carcinoma (IDC) fueron seleccionados como base para esta investigación debido a su simplicidad, baja resolución y disponibilidad pública. Estas características los convierten en candidatos ideales para realizar las primeras pruebas experimentales, sirviendo como una cota mínima para evaluar la factibilidad técnica del enfoque propuesto.

La combinación de ambos datasets también permite explorar una mayor diversidad de escenarios diagnósticos. Por un lado, PCam incluye imágenes histológicas provenientes de múltiples regiones anatómicas y no especifica un tipo particular de cáncer, mientras que IDC se enfoca exclusivamente en tejido mamario y en un subtipo específico de cáncer. Esta diferencia no solo aporta variabilidad morfológica, sino también valor clínico diferenciado. Adicionalmente, PCam presenta una distribución balanceada entre clases (tejido benigno y maligno), mientras que IDC exhibe un desbalance importante, lo que permite evaluar el desempeño de los modelos

en condiciones más desafiantes y representativas del entorno clínico real.

El conjunto PCam está disponible en plataformas ampliamente utilizadas como TensorFlow Datasets [TensorFlow,] y PyTorch [PyTorch,], lo que ha favorecido su adopción en la comunidad científica. Mientras que IDC está disponible en la plataforma Kaggle [Madabhushi and Lee, 2016]. La disponibilidad pública los posiciona como benchmarks consolidados para validar avances en histopatología computacional.

Descripción del Conjunto de Datos Breast Cancer Histopathological Annotation and Diagnosis (BreCaHAD)

El conjunto de datos *Breast Cancer Histopathological Annotation and Diagnosis* (BreCaHAD) fue publicado por [Aksac et al., 2019] con el propósito de facilitar el desarrollo de herramientas computacionales para la detección y clasificación del cáncer de mama a partir de imágenes histopatológicas digitales. El dataset contiene imágenes obtenidas mediante escáner de cortes histológicos teñidos con hematoxilina y eosina (H&E), y está específicamente diseñado para representar regiones relevantes para el diagnóstico clínico del carcinoma ductal invasivo.

Estructura del Dataset

BreCaHAD está compuesto por 162 imágenes de 512×512 píxeles, extraídas de 162 regiones distintas provenientes de 23 muestras de tejido mamario. Cada imagen fue anotada manualmente por patólogos para identificar la presencia de carcinoma invasivo, resultando en un conjunto binario: tejido normal versus tejido canceroso. La resolución elevada permite observar detalles histológicos como variaciones en la forma, tamaño y densidad nuclear, así como la arquitectura general del tejido. Las imágenes están organizadas por paciente y por tipo de tejido, lo que facilita su uso en tareas de clasificación supervisada y evaluación cualitativa.

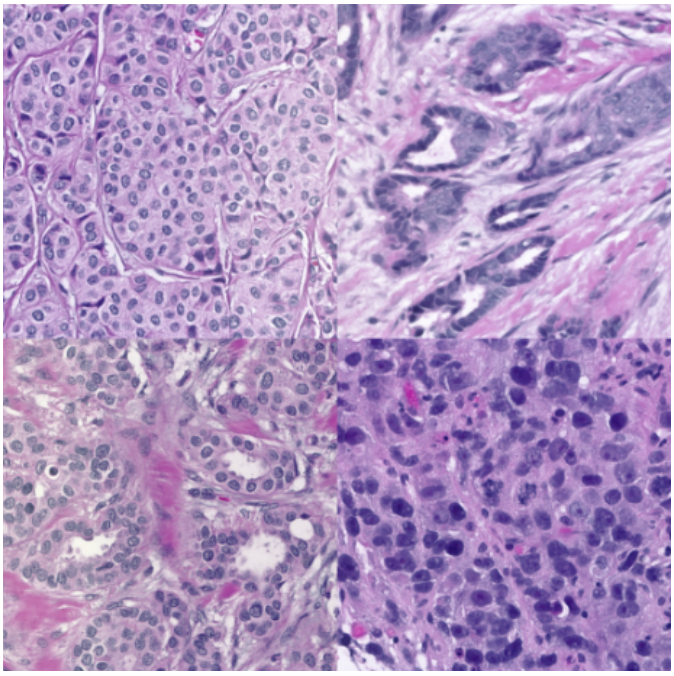


Figura 3.4: **Imágenes extraídas del dataset Breast Cancer Histopathological Annotation and Diagnosis (BreCaHAD).** Las imágenes mostradas corresponden a parches de una resolución de 512×512.

Relevancia para la investigación

El conjunto BreCaHAD fue incorporado en esta investigación debido a dos factores clave: su alta resolución y su tamaño reducido, características que lo convierten en una herramienta ideal para validar la viabilidad técnica de los modelos propuestos en las primeras etapas experimentales. La mayor resolución permite evaluar la capacidad del modelo generativo para representar estructuras morfológicas finas con alto realismo. La idea clave está en conocer los límites superiores del modelo en cuanto a tiempo y complejidad de las imágenes que se pueden utilizar respecto al tiempo de entrenamiento, con los recursos disponibles.

Descripción del Conjunto de Datos NCT-CRC-HE-100K

El NCT-CRC-HE-100K (National Center for Tumor Diseases - Colorectal Cancer - Hematoxylin and Eosin) está compuesto por **100,000 parches** de tejido extraídos de imágenes de microscopía digital de biopsias colorectales. Cada parche tiene una resolución de **256×256 píxeles** en formato ".tiff", con tinción estándar de hematoxilina y eosina (H&E) (ver Figura 3.5), ampliamente utilizada en patología para resaltar estructuras celulares y tisulares.

Los parches fueron obtenidos mediante un proceso de segmentación automatizada de regiones de interés (ROIs) a partir de imágenes de alta resolución (whole-slide images, WSIs). Este enfoque garantiza la extracción de áreas relevantes para el análisis, minimizando la inclusión de artefactos o regiones no informativas. Las anotaciones fueron realizadas por patólogos expertos, asegurando su precisión y coherencia con criterios clínicos establecidos. Adicionalmente, se aplicaron técnicas de normalización de color para mitigar variaciones técnicas asociadas a diferencias en el protocolo de tinción o en el escaneo de las muestras [Kather et al., 2018].

Estructura del Dataset

Este conjunto de datos se divide en **nueve categorías histopatológicas**, que incluyen tejido tumoral (TUM), estroma (STR), músculo liso (MUS), mucosa (MUC), tejido normal (NORM), linfocitos (LYM), tejido adiposo (ADI), fondo (BACK) y restos de tejido (DEB). Para efectos de este trabajo, solo consideramos 2 clases fundamentales para el entrenamiento: tejido tumoral (TUM) y tejido normal (NORM), con 8763 imágenes para cada clase.

Relevancia para la Investigación

Este conjunto de datos es un recurso fundamental en el ámbito de la histopatología computacional, específicamente diseñado para tareas de clasificación de tejidos en cáncer colorrectal (CRC). Fue desarrollado por el National Center for Tumor Diseases (NCT) de Alemania en colaboración con el Instituto de Tecnología de Karlsruhe (KIT), y se ha consolidado como uno de los conjuntos de datos más extensos y referenciados en la literatura científica para el estudio

de imágenes histológicas de CRC.

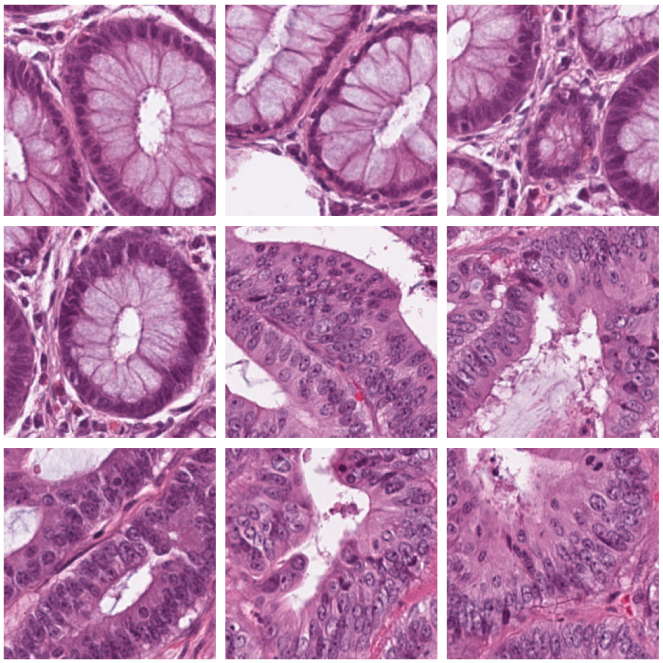


Figura 3.5: **Imágenes extraídas del dataset NCT-CRC-HE-100K.** Las imágenes mostradas corresponden a las clases TUM (tejido canceroso o tumoral) y NORM (tejido normal).

Si bien el conjunto de datos PCam e IDC fueron útiles para las pruebas iniciales, su naturaleza limitada impone restricciones significativas al momento de abordar escenarios clínicos reales. En contextos médicos auténticos, los patólogos no trabajan con imágenes de tan baja resolución como las que ofrecen estos datasets, y además, enfrentan con frecuencia la escasez de datos etiquetados.

Por otra parte, los tejidos histopatológicos presentan una alta complejidad y variabilidad morfológica, incluso dentro de una misma categoría diagnóstica. Imágenes cancerosas extraídas desde distintas regiones del cuerpo pueden mostrar características visuales notablemente distintas entre sí, y lo mismo ocurre con las imágenes benignas. A partir de conversaciones sostenidas con un experto clínico, se concluyó que lo más apropiado sería trabajar con un conjunto de datos

que esté específicamente acotado tanto en términos de la patología analizada como de la región anatómica de extracción. Esta recomendación busca asegurar la coherencia del dominio visual del dataset y, por ende, mejorar la validez de los resultados obtenidos.

Un dataset que cumple con estas condiciones es BreCaHAD, pero dados los recursos computacionales disponibles, se hace prácticamente imposible entrenar la arquitectura propuesta en un tiempo prudente.

Es por lo anterior, que el dataset NCT-CRC-HE-100K es el elegido para realizar la validación del método propuesto en esta tesis junto con todos los experimentos detallados en esta sección. Con él se cumple con una resolución aceptable para que los expertos evalúen las imágenes, una patología específica (cáncer colorectal) y un conjunto relativamente acotado (se utilizará un subconjunto tal que las clases queden balanceadas, con 17.526 imágenes).

La disponibilidad de este conjunto es pública y se encuentra en plataformas como Zenodo [Kather et al., 2018] bajo licencia CC-BY-4.0, lo que ha ampliado su adopción en la comunidad científica y se ha vuelto idóneo para validar avances en histopatología computacional, particularmente en el contexto del cáncer colorrectal.

3.3. Trabajo Previo: Aumento de Datos con StyleGAN2-ADA en Histopatología

La metodología desarrollada en este trabajo previo se fundamentó en el uso de StyleGAN2-ADA como arquitectura generativa para abordar el problema de escasez de datos en la clasificación de imágenes histopatológicas. La selección de esta arquitectura se basó en su capacidad demostrada para producir imágenes de alta calidad incluso cuando se entrena con conjuntos de datos limitados, una característica crucial en el contexto médico donde la obtención de datos etiquetados es costosa y requiere experiencia clínica especializada. StyleGAN2-ADA, como se detalló en la Sección 2, representa una evolución significativa respecto a las arquitecturas GAN tradicionales, incorporando un mecanismo de aumento adaptativo del discriminador que estabiliza el proceso de entrenamiento y mejora la calidad de las imágenes generadas. La arquitectura consta de dos

componentes principales: una red de mapeo que transforma códigos latentes de entrada a un espacio latente intermedio, y una red de síntesis que genera imágenes basadas en estos códigos latentes intermedios. Además, este diseño permite un control más preciso sobre las características de las imágenes generadas, aspecto que resultaría fundamental para el desarrollo posterior del trabajo de explicabilidad contrafactual.

Para llevar a cabo el trabajo propuesto en esta tesis, fue fundamental validar previamente el desempeño de la arquitectura generativa StyleGAN2-ADA en el dominio de imágenes histopatológicas, un contexto particularmente complejo y aún poco explorado en la literatura de modelos generativos avanzados. Específicamente, se requería evaluar si este enfoque era capaz de generar imágenes sintéticas de alta calidad y con un espacio latente suficientemente desentrelazado, a pesar de disponer de un conjunto de datos limitado para el entrenamiento. Asimismo, considerando las restricciones computacionales disponibles (1 GPU NVIDIA TESLA V100 187GB RAM), era indispensable analizar si los tiempos de entrenamiento y generación resultaban viables en relación con las resoluciones de imagen requeridas.

Metodología del Trabajo Previo

Para este trabajo, se utilizaron dos conjuntos de datos histopatológicos descritos anteriormente: PCam (PatchCamelyon) e IDC (Invasive Ductal Carcinoma). La estrategia experimental se diseñó para simular escenarios realistas de limitación de datos, seleccionando aleatoriamente 3 % y 20 % de las imágenes del conjunto de entrenamiento de cada dataset. Esta aproximación permitió evaluar la robustez del método tanto en condiciones de extrema escasez de datos como en escenarios de disponibilidad moderada. Los modelos StyleGAN2-ADA se entrenaron con una configuración específica que incluía resoluciones de 128×128 píxeles para PCam y 64×64 para IDC, 25.000 iteraciones de entrenamiento, tasa de aprendizaje de 0,0025 con optimizador Adam, y parámetros de regularización ajustados para cada dataset.

La implementación del clasificador se basó en una arquitectura ResNet34 preentrenada, modificando las capas finales para adaptarse a la tarea de clasificación binaria específica. La innovación principal del enfoque residió en la estrategia de aumento dinámico, que combina imágenes reales

y sintéticas durante el proceso de entrenamiento según un porcentaje predefinido (r=0,3). Esta estrategia asegura que el modelo se exponga continuamente a nuevos datos sintéticos a lo largo del entrenamiento, potencialmente mejorando sus capacidades de generalización.

Para abordar el problema específico del desbalance de clases presente en el dataset IDC, se implementó una estrategia adicional que consistió en generar datos sintéticos de la clase minoritaria hasta igualar la cantidad de muestras de la clase mayoritaria. Esta aproximación buscaba no solo aumentar la cantidad total de datos disponibles, sino también equilibrar la representación de ambas clases durante el entrenamiento. La evaluación del rendimiento se realizó mediante un conjunto comprensivo de métricas que incluían exactitud, precisión, sensibilidad, F1-score y área bajo la curva ROC (AUC-ROC), proporcionando una visión multidimensional del desempeño del clasificador. Adicionalmente, se monitoreó la Distancia de Fréchet Inception (FID) durante el entrenamiento de los modelos generativos para evaluar la calidad de las imágenes sintéticas producidas.

3.4. Desarrollo del método generativo de imágenes

La presente investigación se basa en una metodología inspirada en dos trabajos recientes y altamente influyentes en el área de explicabilidad mediante modelos generativos: StylEx [Lang et al., 2021] y su posterior aplicación al dominio médico propuesta por Chexplaining in Style [Atad et al., 2022]. La estrategia general consiste en utilizar una arquitectura que incluya un modelo generativo controlable —específicamente, StyleGAN2-ADA— junto con un clasificador y un encoder. Esta combinación permitiría crear un algoritmo para manipular atributos visuales relevantes que influyen en la decisión de un clasificador, generando de esta manera explicaciones contrafactuales visuales sobre imágenes histopatológicas.

El objetivo principal de este enfoque es identificar y modificar atributos latentes que estén estrechamente relacionados con la decisión del clasificador, y visualizar cómo su alteración puede cambiar la predicción del modelo. De este modo, se pretende proporcionar explicaciones visuales plausibles que conecten los factores latentes con decisiones específicas del clasificador.

Arquitectura del método generativo

La arquitectura utilizada está compuesta por tres módulos principales: **un generador condicional** basado en StyleGAN2-ADA, un **clasificador entrenado y congelado**, y un **encoder (o codificador)** que permite proyectar imágenes reales al espacio latente del generador. Esta estructura permite intervenir sobre imágenes reales mediante la manipulación directa del espacio de estilos (StyleSpace), un espacio latente desentrelazado descubierto empíricamente en StyleGAN2 y demostrado como especialmente adecuado para aislar atributos visuales independientes.

A diferencia de una generación tradicional con ruido aleatorio, en este trabajo el entrenamiento se realiza tanto con representaciones latentes provenientes de ruido como con representaciones latentes obtenidas desde imágenes reales mediante el encoder. Esta doble estrategia permite al modelo mantener su capacidad generativa mientras se adapta al dominio visual de interés, asegurando además que las manipulaciones sobre el espacio latente se reflejen con fidelidad en la imagen generada.

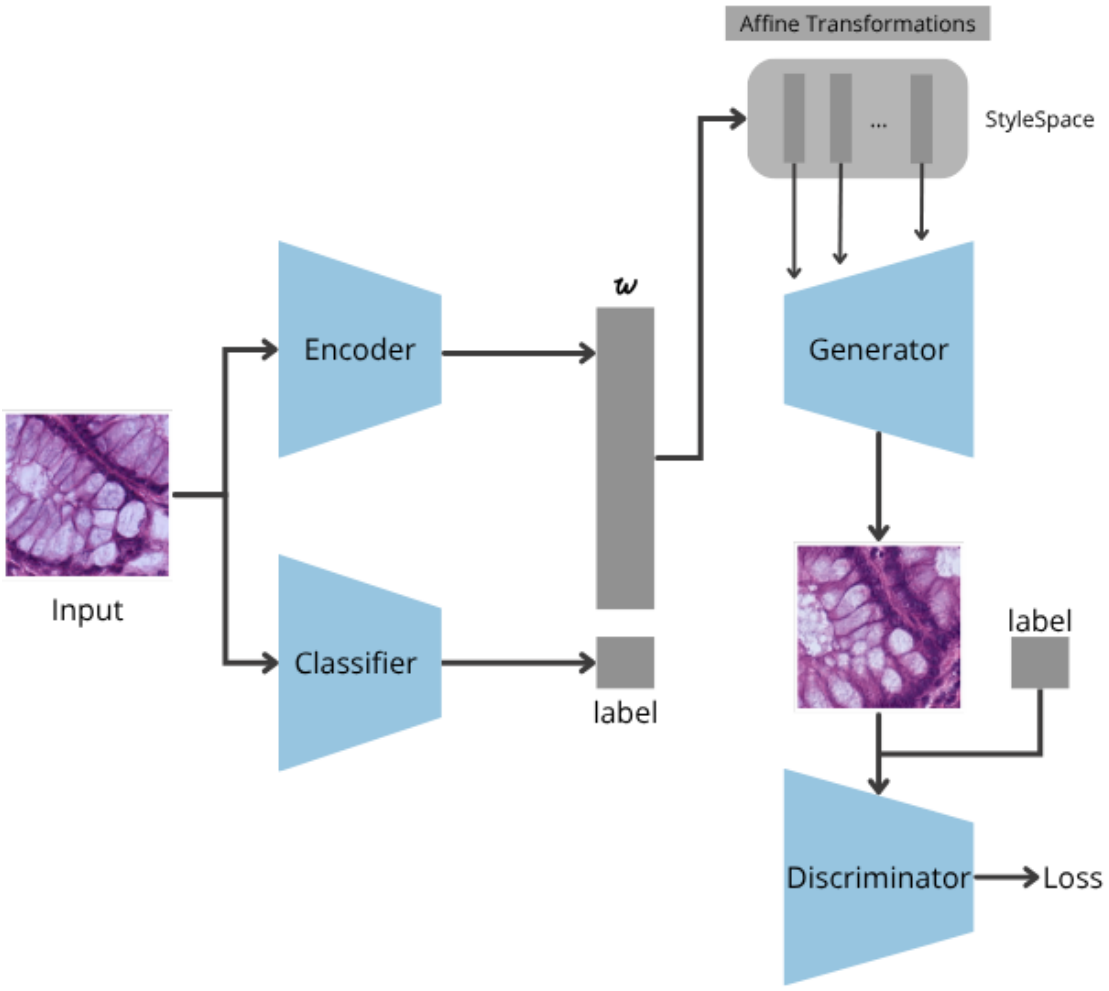


Figura 3.6: **Esquema del modelo generativo propuesto para la generación de imágenes histopatológicas realistas y con un espacio latente desentrelazado.** A partir de una imagen real de entrada, un codificador extrae el vector latente w, el cual es una representación latente de la imagen pero con un espacio latente desentrelazado. Paralelamente, un clasificador preentrenado y con los pesos congelados asigna una etiqueta de clase a la imagen. El vector w, junto con la información de clase, se propaga al generador condicional, que sintetiza una nueva imagen mediante la manipulación de parámetros en el espacio StyleSpace. Esta imagen generada se evalúa con un discriminador, cuya salida contribuye a la función de pérdida durante el entrenamiento adversarial. El diseño permite la exploración estructurada del espacio latente y facilita el análisis de transiciones semánticamente significativas entre clases.

Entrenamiento Guiado por el Clasificador y Funciones de Pérdida

Un aspecto metodológico central de esta investigación, inspirado y adaptado de propuestas como StylEx, es la incorporación explícita de un clasificador pre-entrenado (C) dentro del bucle de entrenamiento del modelo generativo. El objetivo es guiar al generador (G) para que produzca imágenes x_{gen} que no solo sean realistas sino que también preserven las características relevantes para la tarea del clasificador, utilizando una imagen real x_{real} y su representación latente w_{real} (obtenida a través de un encoder E) como punto de partida. Esto se logra mediante una combinación de entrenamiento condicional y una función de pérdida ($\mathcal{L}_{total}$) cuidadosamente diseñada. Esta función se compone de varios términos que balancean dos objetivos principales: la fidelidad a la predicción del clasificador original y la fidelidad de la reconstrucción de la imagen.

1. Pérdida Guiada por el Clasificador ($\mathcal{L}_{cls}$): Este término asegura que la imagen generada x_{gen} produzca una salida en el clasificador C que sea coherente con la salida producida por la imagen real x_{real}. Se mide la divergencia entre las distribuciones de probabilidad predichas (logits) por el clasificador para ambas imágenes. Específicamente, se utiliza la divergencia Kullback-Leibler (D_{KL}) entre las salidas log-softmax, calculada como:

$$\mathcal{L}_{cls} = D_{KL}\Big(\log \sigma\big(C(x_{gen})\big) \parallel \log \sigma\big(C(x_{real})\big)\Big) \tag{3.1}$$

donde $\sigma(\cdot)$ representa la función Softmax. Minimizar $\mathcal{L}_{cls}$ fuerza al generador a prestar atención y representar adecuadamente los atributos visuales que C considera importantes para su predicción.

2. Pérdidas de Reconstrucción: Para garantizar que x_{gen} siga siendo una representación fiel de x_{real} (más allá de las modificaciones inducidas por $\mathcal{L}_{cls}$ o por cambios en el espacio latente si se generan contrafactuales), se emplean varias pérdidas de reconstrucción. Estas miden diferentes aspectos de la similitud entre la imagen generada y la real:

- **Pérdida L1 en el Espacio de la Imagen ($\mathcal{L}_{rec\_x}$):** Fomenta la similitud píxel a píxel

midiendo la diferencia absoluta promedio.

$$\mathcal{L}_{rec\_x} = \|x_{gen} - x_{real}\|_1 \tag{3.2}$$

- **Pérdida L1 en el Espacio Latente ($\mathcal{L}_{rec\_w}$):** Asegura que las representaciones latentes $w = E(x)$ obtenidas por el encoder E sean similares para la imagen real y la generada.

$$\mathcal{L}_{rec\_w} = \|E(x_{gen}) - E(x_{real})\|_1 \tag{3.3}$$

- **Pérdida Perceptual LPIPS ($\mathcal{L}_{lpips}$):** Evalúa la similitud perceptual utilizando características extraídas de una red profunda pre-entrenada (AlexNet). Se calcula sobre imágenes normalizadas por su valor máximo ($\|\cdot\|_\infty$):

$$\mathcal{L}_{lpips} = \text{LPIPS}\left(\frac{x_{gen}}{\|x_{gen}\|_\infty}, \frac{x_{real}}{\|x_{real}\|_\infty}\right) \tag{3.4}$$

3. Función de Pérdida Total ($\mathcal{L}_{total}$): La pérdida final que guía el entrenamiento del generador es una suma ponderada que combina la pérdida del clasificador con un término de reconstrucción $\mathcal{L}_{rec}$:

$$\mathcal{L}_{rec} = \lambda_x \mathcal{L}_{rec\_x} + \lambda_w \mathcal{L}_{rec\_w} + \lambda_{lpips} \mathcal{L}_{lpips} \tag{3.5}$$

La función de pérdida total se expresa entonces como:

$$\mathcal{L}_{total} = \lambda_{cls} \mathcal{L}_{cls} + \lambda_{rec} \mathcal{L}_{rec} \tag{3.6}$$

Donde se probaron distintas configuraciones de pesos λ en función de obtener los mejores resultados posibles, en cuanto a la métrica FID y la calidad de las reconstrucciones.

3.5. Desarrollo del método contrafactual

En este trabajo se propone un método para la generación de contrafactuales en imágenes de histopatología mediante un generador condicional basado en StyleGAN2-ADA, un encoder y un clasificador entrenados en el paso anterior (ver Figura 3.7). **La idea principal es generar múltiples versiones de una misma imagen, manteniendo su estructura global, mediante el uso de un vector de ruido fijo en todas las capas**, mientras se explora el espacio latente a través de la variabilidad de las semillas de entrada. Dicho enfoque busca identificar variaciones en atributos relevantes (por ejemplo, textura, coloración, o morfología local) que influyen en la predicción de un clasificador preentrenado.

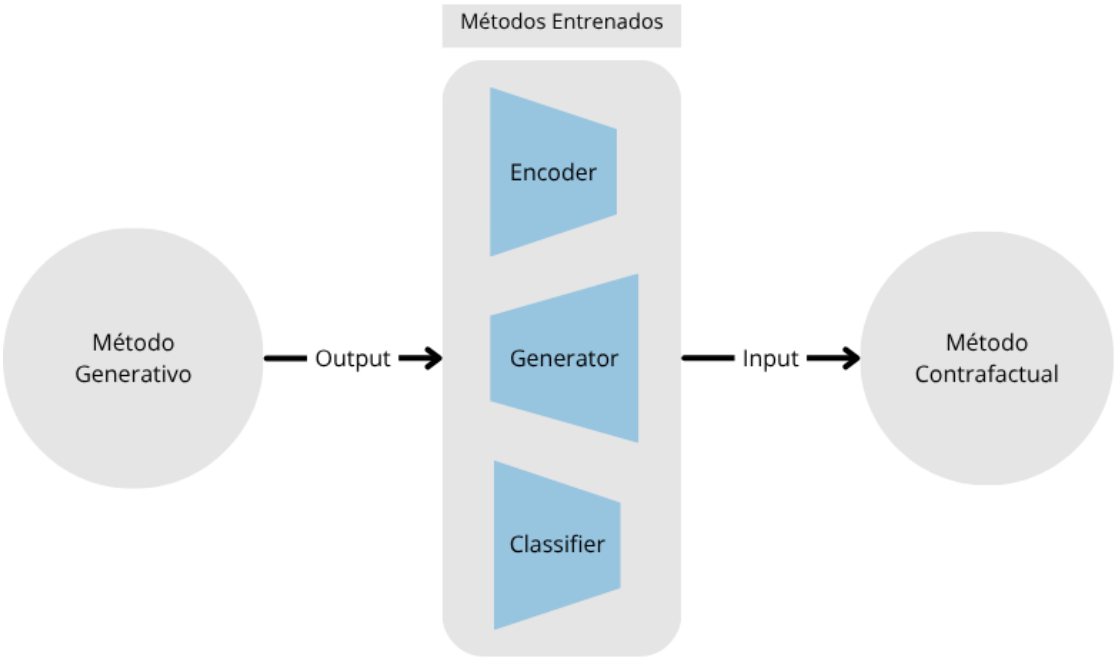


Figura 3.7: **Esquema del método general.** A partir de la arquitectura generativa explicada anteriormente, se obtienen tres modelos entrenados: el generador, el encoder y el clasificador, que posteriormente son utilizados por el método contrafactual propuesto.

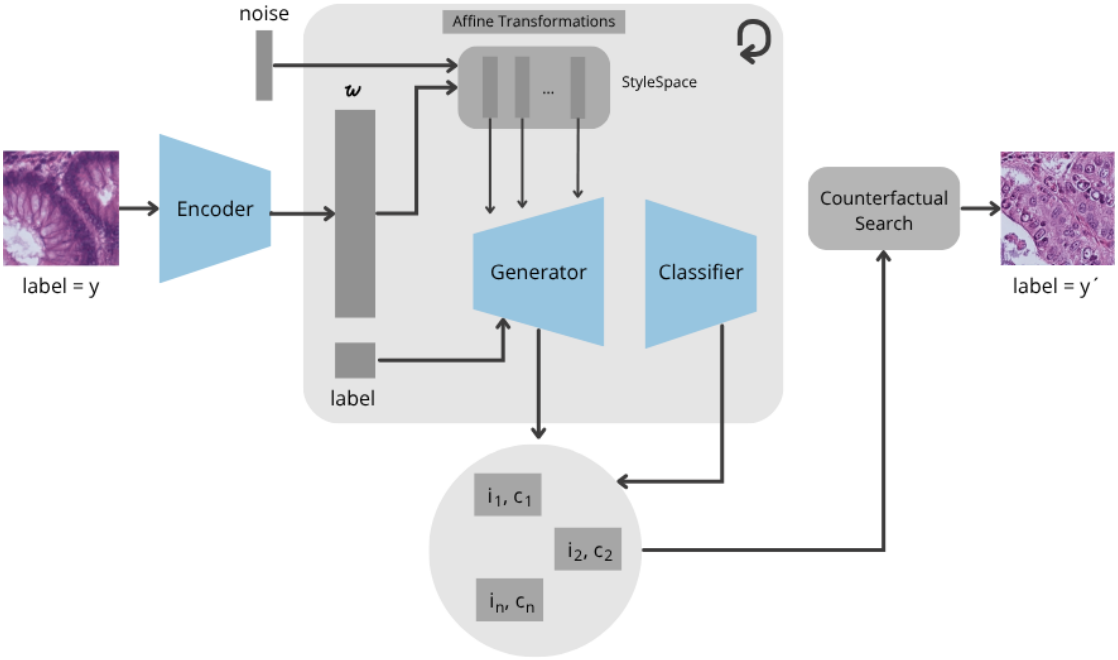


Figura 3.8: **Esquema del método propuesto para la búsqueda de contrafactuales utilizando el generador entrenado basado en StyleGAN2-ADA.** A partir de una imagen de entrada con etiqueta y, se utiliza el codificador entrenado con el modelo generativo (descrito en la Sección 3.6) para producir un vector latente w. Este vector, junto con un conjunto de ruido predefinido fijo transformaciones afines y una etiqueta aleatoria, se inyecta en la red de síntesis (StyleSpace) del generador (3.6), que crea una imagen histopatológica sintética. La imagen generada es evaluada por un clasificador. Este proceso se repite una cantidad N de veces, obteniendo así un conjunto de imágenes con una arquitectura global similar y cuyas diferencias vienen dadas por la etiqueta y variaciones inducidas al transformar la imagen al vector w. A cada imagen i se le asigna un puntaje de clasificación c, que luego es utilizado para elegir y visualizar el contrafactual y' más representativo.

A continuación se presenta el pseudocódigo del algoritmo propuesto para la generación y posterior análisis de imágenes contrafactuales:

Algorithm 1 Generación de Contrafactuales con Ruido Fijo en StyleGAN2-ADA

Require: Generador G, clasificador C, red de transformación f, número de muestras N, dimensión latente d_z, número de capas L

Ensure: Conjunto de imágenes generadas $\{I_i\}$, predicciones $\{p_i\}$, proyecciones latentes U

Inicializar conjuntos vacíos: $Z \leftarrow \varnothing$, $W \leftarrow \varnothing$, $L \leftarrow \varnothing$

Fijar ruido estocástico: para cada $j \in \{1, \ldots, L\}$, muestrear $n^{(j)} \sim \mathcal{N}(0, I)$

for $k = 1$ **to** $\lceil \frac{N}{B} \rceil$ **do**

  Generar batch de vectores $z_i^{(k)} \sim \mathcal{N}(0, I_{d_z})$

  Transformar a espacio w: $w_i^{(k)} = f(z_i^{(k)})$

  Generar etiquetas aleatorias $l_i^{(k)} \sim \text{Uniform}\{0, 1\}$

  Agregar a Z, W, L

end for

Generar imágenes: $I_i \leftarrow G(w_i, l_i, \{n^{(j)}\}_{j=1}^{L})$

Evaluar con el clasificador: $p_i = C(I_i)$

Reducir dimensionalidad: $U \leftarrow \text{UMAP}(W)$

return $\{I_i\}, \{p_i\}, U$

Este método se apoya en tres componentes principales:

- **Generador condicional** G**:** el cual toma como entrada un vector latente intermedio w, una etiqueta condicional l, y un conjunto fijo de tensores de ruido n para producir una imagen sintética I.
- **Encoder** E**:** el cual toma como entrada una imagen real del dataset y la transforma a un vector latente intermedio w. Este componente puede ser o no parte del método, pues también se pueden generar vectores aleatorios w a través del generador. La diferencia está en que al utilizar este componente, los contrafactuales generados serán sobre imágenes reales en vez de sintéticas.

- **Clasificador C:** la red neuronal convolucional entrenada (DenseNet121) utilizada para asignar una probabilidad de pertenencia a una clase binaria (benigno o maligno) a cada imagen generada.

3.5.1. Explicación del Proceso

1. **Inicialización**: Se generan de manera previa los tensores de ruido $\{n^{(j)}\}_{j=1}^{L}$, uno por cada capa que admite inyección de ruido en el generador. Esta inyección se mantiene constante en todas las generaciones, con el fin de preservar los patrones de bajo nivel como la textura de fondo o la forma global de los tejidos, facilitando la interpretación de los cambios inducidos únicamente por las variaciones en w y l.
2. **Exploración del espacio latente**: Se generan N muestras latentes z_i (en B batches o lotes), que se transforman en estilos intermedios w_i mediante la red de transformación f. Se combinan con etiquetas aleatorias l_i para forzar al generador a sintetizar versiones tanto benignas como malignas del mismo estilo base.
3. **Síntesis de imágenes**: Con los estilos w_i, las etiquetas l_i y el conjunto fijo de ruido, se generan imágenes I_i.
4. **Evaluación y análisis**: Cada imagen se clasifica con el modelo C para observar cómo la modificación semántica (provocada por l_i o w_i) afecta la decisión del clasificador. El conjunto de vectores w_i se proyecta a dos dimensiones con UMAP para visualizar cómo se estructuran las representaciones latentes en función de la salida del clasificador.

3.5.2. Importancia del Enfoque

Este enfoque permite generar un **conjunto diverso de contrafactuales altamente interpretables**, ya que mantiene la coherencia estructural de las imágenes y permite observar cómo cambios sutiles en los estilos (atributos semánticos) afectan la clasificación. La fijación del ruido actúa como mecanismo de control que permite atribuir los cambios de clase únicamente a la variación semántica y no a factores aleatorios o irrelevantes.

3.5.3. Interpolación Latente para Visualización de Contrafactuales

Además del muestreo aleatorio de imágenes condicionadas, se propone una estrategia complementaria basada en interpolaciones latentes. Dado un vector latente w_{orig} correspondiente a una imagen de clase $c \in \{0, 1\}$, se selecciona un segundo vector w_{cf} correspondiente a una imagen generada por el mismo estilo w pero con la clase opuesta $\bar{c}$.

Para visualizar de forma progresiva la transición entre ambas clases, se realiza una interpolación lineal entre los dos vectores latentes:

$$w^{(\alpha)} = (1 - \alpha)w_{\text{orig}} + \alpha w_{\text{cf}}, \quad \alpha \in [0, 1] \tag{3.7}$$

Cada vector $w^{(\alpha)}$ se combina con la etiqueta condicional $l = \bar{c}$ y los mapas de ruido fijos $\{n^{(j)}\}_{j=1}^{L}$ para generar una imagen $I^{(\alpha)}$:

$$I^{(\alpha)} = G(w^{(\alpha)}, \bar{c}, \{n^{(j)}\}) \tag{3.8}$$

Estas imágenes interpoladas se evalúan con el clasificador C para obtener su puntaje de clase:

$$p^{(\alpha)} = C(I^{(\alpha)}) \tag{3.9}$$

Esta interpolación permite:

- Visualizar de manera continua y suave los cambios morfológicos a lo largo de una transición semántica entre clases.
- Visualizar de manera gradual los cambios en la imagen que impactan en el puntaje de clasificación.
- Visualizar la frontera de decisión del clasificador.
- Analizar qué transformaciones visuales específicas inducen cambios significativos en la clasificación, lo cual es crucial para el entendimiento y la validación de modelos en entornos médicos sensibles.

Pseudocódigo de Interpolación Contrafactual

Algorithm 2 Interpolación en el Espacio Latente entre Clases Contrarias

Require: Vectores $w_{\text{orig}}, w_{\text{cf}}$, clase original c, generador G, clasificador C, pasos K, mapas de ruido $\{n^{(j)}\}$

for $k = 0$ **to** K **do**
  $\alpha \leftarrow \frac{k}{K}$
  $w^{(\alpha)} \leftarrow (1-\alpha)w_{\text{orig}} + \alpha w_{\text{cf}}$
  $I^{(\alpha)} \leftarrow G(w^{(\alpha)}, \bar{c}, \{n^{(j)}\})$
  $p^{(\alpha)} \leftarrow C(I^{(\alpha)})$
end for
return Secuencia $\{(I^{(\alpha)}, p^{(\alpha)})\}_{k=0}^{K}$

Análisis Visual

Se presenta la secuencia de imágenes $\{I^{(\alpha)}\}$ junto con sus respectivos valores de clasificación $p^{(\alpha)}$ (un ejemplo de secuencia de intepolación se puede ver en la Figura 3.9). A esta interpolación podemos llamarle también trayectoria contrafactual. Esta visualización revela los puntajes de clasificación de cada imagen a medida que esta transita de una clase a otra, permitiendo identificar paso a paso aquellos cambios que van produciendo el cambio de clasificación.

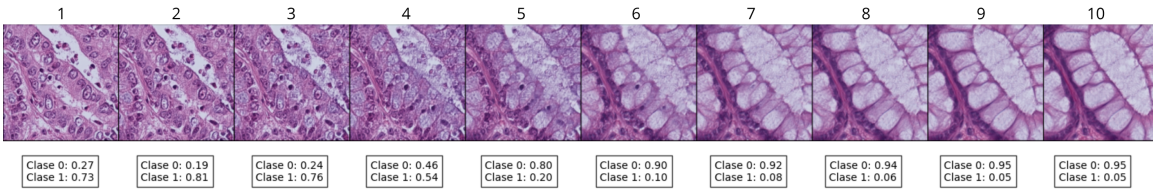


Figura 3.9: **Ejemplo de secuencia de interpolación desde clase cancerosa a benigna**.

3.6. Exploración y análisis del espacio latente

En este apartado se detalla el proceso de exploración del espacio latente con el objetivo de identificar patrones semánticos y característicos que puedan relacionarse con la decisión del clasificador. La exploración se realiza bajo dos configuraciones distintas: (i) utilizando un vector de ruido fijo en todas las capas del generador y (ii) variando el ruido de forma aleatoria en cada generación. Esta estrategia permite evaluar la influencia del ruido en la coherencia estructural y en la diversidad de las imágenes generadas, así como la capacidad del espacio latente de representar diferencias significativas entre clases.

Motivación y Objetivos

Para evaluar la capacidad de separación de clases en el espacio latente, se proyectan tanto los vectores iniciales z como los vectores intermedios w (resultado de la transformación $f(z)$) en un espacio bidimensional mediante técnicas de reducción de dimensionalidad (UMAP). Esto permite visualizar si, y en qué medida, las muestras de distintas clases se separan en el espacio latente, contribuyendo a la interpretación semántica y a la validación del proceso de *disentanglement*.

El espacio latente generado por la red de transformación f posee propiedades de desentrelazamiento o *disentanglement* que, en principio, facilitan la interpretación de atributos semánticos relevantes. Sin embargo, la influencia de la inyección de ruido en las distintas capas puede condicionar tanto la consistencia de la estructura global como la representación de detalles locales. Por ello, se plantea la siguiente doble configuración:

- **Ruido fijo:** Al utilizar un conjunto único de tensores de ruido $n = \{n^{(j)}\}_{j=1}^{L}$ para todas las muestras, se preserva la estructura global de las imágenes. Esto permite analizar cómo la variabilidad en el vector latente w se traduce en cambios en atributos semánticos de nivel bajo a medio sin que la estructura global se vean afectada. Se utiliza para la generación contrafactual.

- **Ruido variable:** En esta configuración se genera un nuevo tensor de ruido para cada imagen, lo que introduce variaciones tanto en la estructura global como en los detalles. Esta opción favorece la diversidad en la síntesis a la vez que permite analizar los diferentes tipos de estructuras codificadas en el espacio latente.

En esta exploración se construye un *diagrama de densidad* sobre el espacio latente, así se pueden identificar las regiones de alta densidad por cada clase. A partir de ello se extraen y visualizan las imágenes correspondientes a estos puntos (o clusters), permitiendo establecer vínculos entre la posición en el espacio y las características visuales relevantes (por ejemplo, la morfología del tejido o la distribución de núcleos) que, en última instancia, puedan ser consideradas representativas para la interpretación de la predicción.

Procedimiento y Análisis

El procedimiento experimental se resume en los siguientes pasos:

1. **Generación de Muestras en el Espacio Latente:** Se generan N vectores $z_i \sim \mathcal{N}(0, I)$ y se transforman a $w_i = f(z_i)$ utilizando el generador condicional. Para cada vector w_i, se asocian dos configuraciones de síntesis:
   - Con **ruido fijo**: se utiliza un conjunto predefinido de tensores n inyectados en todas las capas.
   - Con **ruido variable**: se genera un nuevo conjunto de tensores de ruido para cada muestra.
2. **Síntesis de Imágenes:** Para cada configuración se generan las imágenes I_i a partir de w_i y de la etiqueta condicional l_i, utilizando el mismo procedimiento de la sección anterior. Estas imágenes son evaluadas con el clasificador C para obtener puntajes de predicción que se emplearán en el análisis.
3. **Construcción del Diagrama de Densidad:** Se aplica una reducción de dimensionalidad al conjunto de vectores $w = \{w_i\}_{i=1}^{N}$ mediante UMAP [McInnes et al., 2020], obteniendo una proyección $U \subset \mathbb{R}^2$. Sobre esta proyección se calcula la densidad de puntos mediante

estimación de densidad KDE (Kernel Density Estimation) para determinar las regiones de alta concentración de muestras.

4. **Selección y Visualización de Muestras Representativas:** Se identifican los puntos de mayor densidad en U para cada clase y se recuperan las imágenes correspondientes a dichos vectores latentes. De esta forma, se genera un compendio visual que permite analizar si existe algún patrón o característica representativa asociado a cada región del espacio latente y, consecuentemente, a la clase a la cual pertenecen.

Método de Referencia: Chexplaining in Style

Para establecer una línea base comparativa con el método propuesto en esta investigación, se implementó el enfoque descrito en "Chexplaining in Style'" [Atad et al., 2022], el cual representa una adaptación del método StylEx [Lang et al., 2021] específicamente diseñada para el dominio médico. Esta implementación permite evaluar de manera objetiva las ventajas y limitaciones de la metodología desarrollada en el presente trabajo.

Fundamentos del Método Chexplaining in Style

Chexplaining in Style constituye una extensión del framework StylEx al ámbito de la imagenología médica, específicamente diseñado para generar explicaciones contrafactuales en radiografías de tórax. El método se basa en la premisa de que las representaciones latentes de StyleGAN2 pueden manipularse de manera controlada para identificar y visualizar los factores que influyen en las decisiones de clasificación médica.

La arquitectura fundamental del método incluye las mismas tres componentes descritas en 3.4. El generador StyleGAN2 se entrena en el dominio específico de imágenes médicas para generar muestras sintéticas de alta calidad que preserven las características morfológicas relevantes del tejido. El encoder de proyección actúa como una red neuronal que mapea imágenes reales al espacio latente del generador, permitiendo la manipulación posterior de estas representaciones. Finalmente, el clasificador preentrenado proporciona las predicciones de referencia y guía el

proceso de generación de explicaciones.

Las principales diferencias de este método con la implementación propuesta en esta tesis radican en la incorporación de StyleGAN2-ADA (en lugar de StyleGAN2), el cual introduce una estrategia de aumento de datos adaptativa que mejora la generalización del modelo, especialmente en escenarios con poca disponibilidad de datos (no es lo mismo entrenar un modelo con 100mil imágenes histopatológicas que con 100mil imágenes de otro dominio, ya que la complegidad de estas imágenes hace que sea más difícil la convergencia de los modelos. Además, para que un experto pueda evaluar las imágenes en cuanto a diagnóstico, se requiere una mayor resolución, es decir, los 100mil parches de imagen de 50 × 50 se transforman en mil parches de 512 × 512) o alta variabilidad morfológica, como es común en imágenes histopatológicas. Además, se realizaron adaptaciones específicas en el entrenamiento para ajustarse a los datasets histopatológicos utilizados, así como modificaciones en la configuración de las funciones de pérdida para mejorar la calidad de las imágenes generadas.

Proceso de Generación de Explicaciones

El método Chexplaining in Style opera mediante un proceso de optimización en el espacio latente que busca encontrar la mínima modificación necesaria para cambiar la predicción del clasificador. A diferencia de métodos que optimizan directamente en el espacio latente completo, este enfoque utiliza un proceso iterativo basado en el Análisis de Componentes Principales (PCA) [Jolliffe and Cadima, 2016] para identificar de manera sistemática las direcciones más relevantes para la generación de contrafactuales.

El algoritmo comienza generando un conjunto diverso de vectores latentes w_i mediante muestreo aleatorio del espacio latente de StyleGAN2, con el objetivo de cubrir diferentes regiones del espacio de características y capturar la variabilidad inherente del dominio médico. Cada uno de estos vectores latentes se utiliza posteriormente para generar una imagen sintética correspondiente $x_i = G(w_i)$ utilizando el generador StyleGAN2 que ha sido específicamente entrenado en el dominio médico de interés. Una vez generado este conjunto de imágenes sintéticas, cada una de ellas se evalúa utilizando el clasificador médico preentrenado para obtener las prediccio-

nes correspondientes $y_i = C(x_i)$. Este paso es fundamental ya que establece la relación directa entre las representaciones latentes y las decisiones de clasificación, proporcionando la información necesaria para identificar qué regiones del espacio latente están asociadas con diferentes predicciones médicas.

El método procede entonces aplicando PCA sobre el conjunto completo de vectores latentes w_i. Esta técnica estadística identifica las direcciones de máxima varianza en el espacio latente, revelando así las dimensiones que capturan la mayor parte de la variabilidad presente en las representaciones. Las componentes principales resultantes representan direcciones ortogonales que, en principio, corresponden a variaciones semánticamente coherentes en las imágenes generadas.

A partir de este análisis, se seleccionan las k primeras componentes principales que explican un porcentaje significativo de la varianza total, típicamente entre 80 % y 90 %. La premisa fundamental es que estas direcciones principales contienen la información semántica más relevante para la diferenciación entre clases médicas, mientras que las componentes de menor varianza corresponden principalmente a variaciones aleatorias o ruido.

Para validar esta suposición, el método evalúa la relevancia semántica de cada componente principal seleccionada mediante la manipulación controlada de vectores latentes a lo largo de cada dirección identificada. Esto se realiza generando imágenes con diferentes magnitudes de desplazamiento en cada componente y evaluando cómo estos cambios afectan las predicciones del clasificador médico. Las componentes que inducen cambios más significativos y consistentes en las predicciones se consideran más relevantes para la tarea de explicabilidad.

El proceso adopta un enfoque iterativo donde, basándose en los resultados obtenidos en cada ciclo, se refina el muestreo del espacio latente para concentrarse en regiones que muestran mayor sensibilidad a los cambios de clasificación. Este refinamiento progresivo permite una exploración más eficiente del espacio latente y una identificación más precisa de las direcciones semánticamente relevantes.

Una vez identificadas las direcciones relevantes mediante PCA, el proceso de generación de contrafactuales se formula como:

$$\mathbf{w}' = \mathbf{w} + \sum_{i=1}^{k} \alpha_i \cdot \mathbf{v}_i \tag{3.10}$$

donde:

- $\mathbf{w}$ es la representación latente de la imagen original,
- $\mathbf{v}_i$ son las componentes principales seleccionadas,
- α_i son los coeficientes de manipulación para cada componente,
- k es el número de componentes principales utilizadas.

Los coeficientes α_i se optimizan mediante el siguiente objetivo:

$$\{\alpha_i\} = \arg\min_{\{\alpha_i\}} \|\boldsymbol{\alpha}\|_2 + \lambda \cdot \mathcal{L}_{\text{cls}}\left(C\left(G\left(\mathbf{w} + \sum_i \alpha_i \cdot \mathbf{v}_i\right)\right), y_{\text{target}}\right) \tag{3.11}$$

A diferencia de aplicaciones en otros dominios, en Chexplaining in Style dicen incorporar consideraciones específicas para imágenes médicas que reflejan la naturaleza crítica de este contexto (un ejemplo se muestra en la Figura 3.10). El método incluye restricciones adicionales para asegurar que las modificaciones no alteren estructuras anatómicas fundamentales, manteniendo así la plausibilidad médica de las imágenes generadas. Adicionalmente, se implementan mecanismos para limitar la magnitud de las modificaciones, evitando transformaciones excesivas que podrían resultar en imágenes médicamente irreales. El método también incluye una evaluación por parte de expertos médicos, reconociendo la importancia fundamental de la validación clínica en aplicaciones de salud.

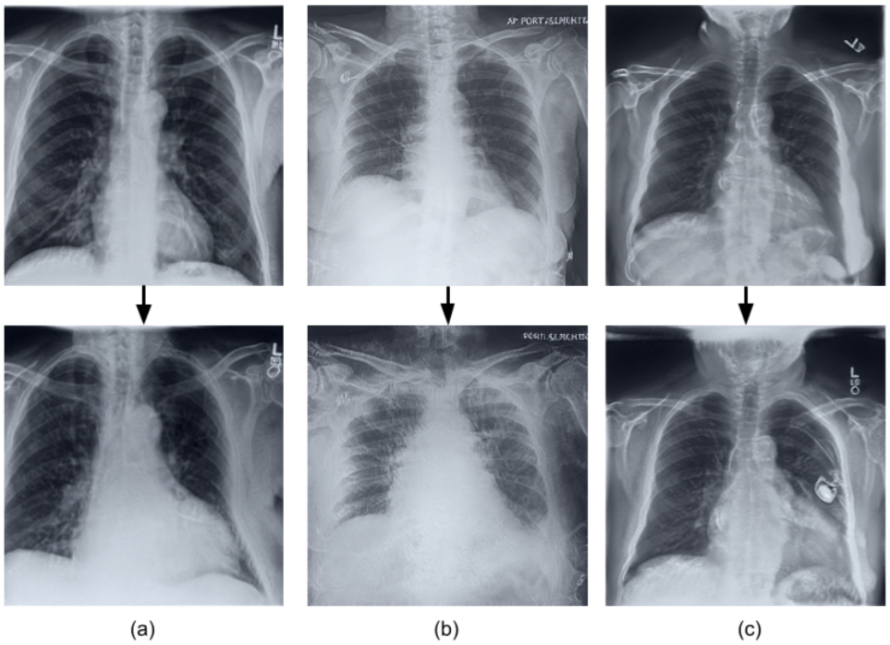


Figura 3.10: **Comparaciones entre imágenes originales (primera fila) y contrafactuales generadas (segunda fila) por Chexplaining in Style.** Las modificaciones reflejan características clínicas relevantes para cada diagnóstico, como cardiomegalia (a), derrame pleural (b) y presencia de marcapasos (c).

Justificación para la Comparación

La implementación de Chexplaining in Style como método de referencia se justifica por múltiples razones que fortalecen la validez de la comparación experimental. En primer lugar, constituye uno de los pocos métodos específicamente adaptados para explicabilidad en imágenes médicas utilizando modelos generativos, lo que lo convierte en un punto de referencia directo y relevante para el dominio de aplicación. Además, será valioso conocer los resultados del método sobre el dominio histopatológico, que como ya se ha explicado anteriormente, presenta múltiples desafíos y complejidades no abordadas hasta ahora.

Para una justa comparación se utilizará la misma estructura generativa entrenada, es decir, el punto de partida será el método generativo StyleGAN2-ADA, el encoder y el clasificador ya

entrenados sobre el dataset NCT-CRC-HE-100K. Esto permitirá una comparación más directa de las estrategias de manipulación latente sin que las diferencias arquitectónicas fundamentales distorsionen los resultados.

Esta implementación comparativa permite evaluar objetivamente las contribuciones específicas del método propuesto, particularmente en términos de eficiencia computacional, calidad de las explicaciones contrafactuales generadas y aplicabilidad al dominio de histopatología.

3.7. Software de evaluación con expertos: Validación cualitativa

Se desarrolló un software para *validar cualitativamente* la utilidad e interpretabilidad del marco implementado, asegurando su coherencia con el conocimiento médico y la calidad visual de los resultados. En particular, se abordan los siguientes puntos:

1. **Coherencia con el conocimiento médico:** Validar si las características visuales en las imágenes contrafactuales generadas coinciden con el conocimiento experto.
2. **Realismo y consistencia visual:** Verificar si el realismo visual y los patrones biológicos se mantienen al generar imágenes sintéticas.
3. **Utilidad en la toma de decisiones:** Evaluar si las explicaciones contrafactuales refuerzan o facilitan el proceso de diagnóstico.

3.7.1. Diseño del Experimento

Para llevar a cabo el experimento se desarrolló una plataforma web construida con React, TypeScript y Supabase como sistema de base de datos. Esta herramienta de acceso abierto (disponible en histoXplain) se compartió y difundió entre expertos de distintas áreas.

La aplicación se organiza en **cuatro etapas**, diseñadas para evaluar los tres objetivos definidos previamente. Se pueden ver algunos pantallazos de cada etapa en el Anexo 5.

Etapa 1: Evaluación de Imágenes Sintéticas

Se presenta a cada experto, de manera *aleatoria*, un conjunto de **20 imágenes**:

- 50 % imágenes reales (5 imágenes de cada clase),
- 50 % imágenes sintéticas generadas por el modelo (5 imágenes de cada clase).

Por cada imagen mostrada:

1. El experto indica si la imagen es *real*, *sintética* o *indistinguible*.
2. La respuesta se registra y se avanza a la siguiente imagen hasta completar todas las imágenes.

Esta primera parte del experimento nos permite conocer si las imágenes generadas por el modelo para ambas clases presentan un nivel de realismo indistinguible respecto a las imágenes reales.

Etapa 2: Generación y evaluación de Imágenes Contrafactuales

Se presenta a cada experto 10 visualizaciones, cada una con una imagen inicial junto a su imagen contrafactual correspondiente. Adicionalmente, se le muestra una secuencia de imágenes a modo de interpolación entre la imagen inicial y la imagen contrafactual. Esto permite que el usuario visualice imágenes intermedias entre ambas imágenes.

Junto a la colaboración con una experta, se ha predefinido un conjunto de 5 características visuales que se espera que estén presentes en una imagen histopatológica de tipo cancerosa, las 3 primeras son características primarias y las 2 restantes son características secundarias. El objetivo de esta evaluación es determinar si la imagen contrafactual generada contiene o no cada una de estas características:

1. **Arquitectura glandular alterada**: Muestra una pérdida de la arquitectura normal, con formación de glándulas irregulares, desorganizadas o estructuras sólidas sin una organización glandular clara.
2. **Pleomorfismo celular y nuclear**: Se observa variabilidad en el tamaño y forma de las

células y sus núcleos (pleomorfismo), núcleos hipercromáticos y aumento de la relación núcleo/citoplasma.

3. **Incremento de la actividad mitótica**: Hay un aumento de figuras mitóticas en diversas áreas del tumor, indicando una proliferación celular descontrolada

4. **Invasión del estroma y otras capas**: Las células tumorales invaden la submucosa, la muscular propia y pueden afectar estructuras adyacentes, reflejando la capacidad invasiva de cáncer.

5. **Producción excesiva de moco**: Algunos adenocarcinomas producen cantidades excesivas de moco, formando lagunas mucosas en el tejido tumoral. En eosina se puede ver transparencia o rosado claro.

Esta segunda parte del experimento nos permite conocer si las imágenes contrafactuales generadas por el modelo contienen o no características patológicas asociadas a su clase.

Etapa 3: Evaluación de Consistencia Visual en Imágenes Reconstruidas

Se presentan al experto 10 visualizaciones, cada una con una imagen real obtenida directamente desde el dataset junto con una imagen recontruida generada por el modelo al pasar esa imagen real por el encoder. El objetivo es evaluar si la imagen reconstruida sigue siendo realista y mantiene el patrón biológico de la imagen real. Para eso se le realizan al experto las siguientes preguntas en cada visualización:

1. ¿El patrón biológico de la imagen original se mantiene o se pierde en la reconstrucción? *(Opciones: "Se mantiene", "Se pierde parcialmente", "Se pierde por completo").*

2. ¿La imagen reconstruida sigue siendo realista visualmente? *(Escala de 1 a 5, refiriéndose a poco realista y muy realista respectivamente).*

Etapa 4: Utilidad del mecanismo

En esta última etapa, cada experto responde a las siguientes preguntas utilizando una escala de Likert (1-5) [Likert, 1932]:

1. ¿La visualización de los contrafactuales es útil para reforzar/apoyar tu decisión diagnóstica?
2. ¿Crees que una herramienta de este estilo sería útil para aquellos casos difíciles de reconocer al ojo humano (casos equívocos o intermedios)?

El objetivo es conocer si los expertos encuentran algún valor en la visualización de los contrafactuales para el diagnóstico médico.

Evaluación Integral del Modelo: Combinación de Métricas

En el contexto de este experimento, es fundamental destacar que **ninguna métrica individual es suficiente para evaluar completamente la calidad y utilidad de un modelo generativo**. Cada métrica o prueba aborda aspectos específicos del desempeño del modelo, pero su interpretación aislada puede llevar a conclusiones parciales o incompletas. Por esta razón, se ha diseñado un enfoque integral que combina múltiples métodos de evaluación para garantizar una validación robusta y coherente con los objetivos planteados.

1. Test de Turing Visual

La primera etapa del experimento incluye la realización de un **Test de Turing Visual (Visual Turing Test, VTT)**, en el cual expertos deben distinguir entre imágenes reales, sintéticas e indistinguibles. Este test permite evaluar el **realismo perceptual** de las imágenes generadas, es decir, si las imágenes sintéticas son lo suficientemente realistas como para engañar a un observador humano entrenado [Chuquicusma et al., 2018]. Un valor de precisión global (accuracy) perfecto sería de 50 %, esto quiere decir que los expertos son incapaces de detectar con certeza absoluta si las imágenes generadas son reales o sintéticas.

La precisión global se define como

$$\text{Accuracy} = \frac{\#\{\text{Reales correctamente identificadas}\} + \#\{\text{Sintéticas correctamente identificadas}\}}{\#\{\text{Total de respuestas}\}}$$

En donde, si el experto lograra identificar a la perfección todas las reales y sintéticas, significa que "se nota la diferencia" entre las imágenes reales y las generadas por el modelo (Accuracy = 1). Por otro lado, si el experto no lograra identificiar ninguna imagen real y tampoco ninguna sintética (Accuracy = 0), podría significar que el usuario ignora completamente el dominio o que ha confundido las reglas, porque es muy poco probable incluso haciendo esto al azar. El caso ideal sería en donde el experto no sea capaz de distinguir las reales de las sintéticas fallando el aproximadamente el 50 % de los casos (Accuracy = 0,5).

Aunque el VTT es útil para medir el realismo visual, no garantiza que las imágenes sean útiles desde un punto de vista médico ni que preserven características biológicas relevantes. Por ello, su uso se complementa con otras métricas y evaluaciones.

2. FID (Frechet Inception Distance)

Para complementar el análisis subjetivo del VTT, se utiliza la métrica técnica **FID (Frechet Inception Distance**, ver Sección 2.5), que cuantifica la similitud estadística entre las distribuciones de imágenes reales y sintéticas. A diferencia del VTT, el FID no depende de la percepción humana y proporciona una medida objetiva de la calidad técnica de las imágenes generadas. Sin embargo, al igual que el VTT, el FID no asegura que las imágenes sean interpretables o útiles en un contexto clínico, lo que subraya la necesidad de incorporar evaluaciones adicionales.

3. Índice Compuesto de Plausibilidad Visual S

Con el fin de cuantificar de forma unificada el grado de "cancerosidad" percibida en cada contrafactual en la Etapa 2 (3.7.1), se definió un *Índice Compuesto de Plausibilidad* S basado en la presencia o ausencia de cinco características visuales relevantes (tres primarias y dos secundarias).

Las características evaluadas son las 5 especificadas en 3.7.1, donde cada característica $i \in \{1, \ldots, 5\}$ es valorada por el experto como $\mathrm{presente}$ (1) o $\mathrm{ausente}$ (0).

Para reflejar la mayor relevancia diagnóstica de las tres características primarias, se asigna un peso diferenciado para rasgos primarios y secundarios.

Cálculo del índice Sea $p_i \in \{0, 1\}$ la respuesta binaria para la característica i. Se asignan pesos diferenciados: $w_i = 1$ para las características primarias ($i = 1, 2, 3$) y $w_i = \frac{1}{2}$ para las características secundarias ($i = 4, 5$). Se define

$$S = \frac{\sum_{i=1}^{3} w_i\, p_i + \sum_{i=4}^{5} w_i\, p_i}{\sum_{i=1}^{3} w_i + \sum_{i=4}^{5} w_i} \times 100 \quad (\,\%)$$

donde $\sum_{i=1}^{3} w_i + \sum_{i=4}^{5} w_i = 3 \cdot 1 + 2 \cdot \frac{1}{2} = 4$. El índice S oscila entre $0\,\%$ (ningún rasgo relevante presente) y $100\,\%$ (todas las características relevantes presentes).

Interpretación y uso

- Un valor de S cercano a $100\,\%$ indica que el contrafactual exhibe la mayoría de los rasgos patológicos esperados en un tejido canceroso.
- Un valor de S próximo a $0\,\%$ sugiere que el contrafactual carece de los signos visuales asociados a malignidad.
- Este índice permite, a su vez, establecer un *umbral de decisión* τ (por ejemplo $\tau = 50\,\%$) para clasificar automáticamente cada imagen como “cancerosa” ($S \geq \tau$) o “benigna” ($S < \tau$), de esta manera podemos comparar este índice con la etiqueta original de la imagen.

Con este procedimiento garantizamos una métrica objetiva y reproducible de plausibilidad visual, que sintetiza en un único valor la evaluación experta de múltiples atributos morfológicos.

4. Encuestas de Evaluación de Características Relevantes

Junto con el Índice Compuesto S, se realizan encuestas detalladas en las **etapas 3 y 4** del experimento para evaluar aspectos específicos de las imágenes generadas. En la **etapa 3**, los expertos evalúan la **consistencia visual** y el **realismo percibido** de las imágenes reconstruidas mediante preguntas estructuradas. En la **etapa 4**, se analiza la **utilidad del mecanismo** en términos de apoyo al diagnóstico y manejo de casos difíciles. Estas encuestas permiten obtener información cualitativa valiosa sobre la **interpretabilidad** y **aplicación práctica** del modelo, aspectos que no pueden ser capturados por métricas técnicas como el FID.

Justificación de la Combinación de Métodos

La combinación de estas evaluaciones —**Test de Turing Visual**, **FID**, **Índice Compuesto de Plausibilidad S** y **encuestas de características relevantes**— responde a la necesidad de abordar múltiples dimensiones del desempeño del modelo. Mientras que el VTT y el FID se enfocan en el **realismo visual y técnico**, respectivamente, el Índice Compuesto S y las encuestas de las etapas 3 y 4 aseguran que las imágenes generadas cumplan con criterios de **utilidad clínica** e **interpretabilidad**. Este enfoque híbrido permite validar no solo la calidad visual de las imágenes, sino también su capacidad para **reforzar el proceso de diagnóstico** y **aumentar la confianza** de los expertos en las predicciones del modelo [Higaki et al., 2022].

Resultados y Análisis esperados

1. **Etapa 1:**
    - Porcentaje de imágenes clasificadas correctamente como reales, sintéticas o indistinguibles. (Test de Turing Visual)

2. **Etapa 2:**
    - Valor de la métrica compuesta S para conocer si las características que definen a una imagen histopatológica cancerosa se encuentran o no presentes en los contrafactuales cancerosos y benignos generados, respectivamente.
    - Matriz de confusión para conocer la sensibilidad y especificidad de la métrica compuesta S.
3. **Etapa 3:**
    - Porcentaje de imágenes donde el patrón biológico se mantiene.
    - Realismo visual promedio de la reconstrucción (escala de Likert)
4. **Etapa 4:**
    - Porcentaje de respuestas (1-5) en la escala de Likert para las dos preguntas relacionadas a la utilidad del método en la práctica.

Capítulo 4

Resultados y Discusión

4.1. Resultados del Trabajo Previo

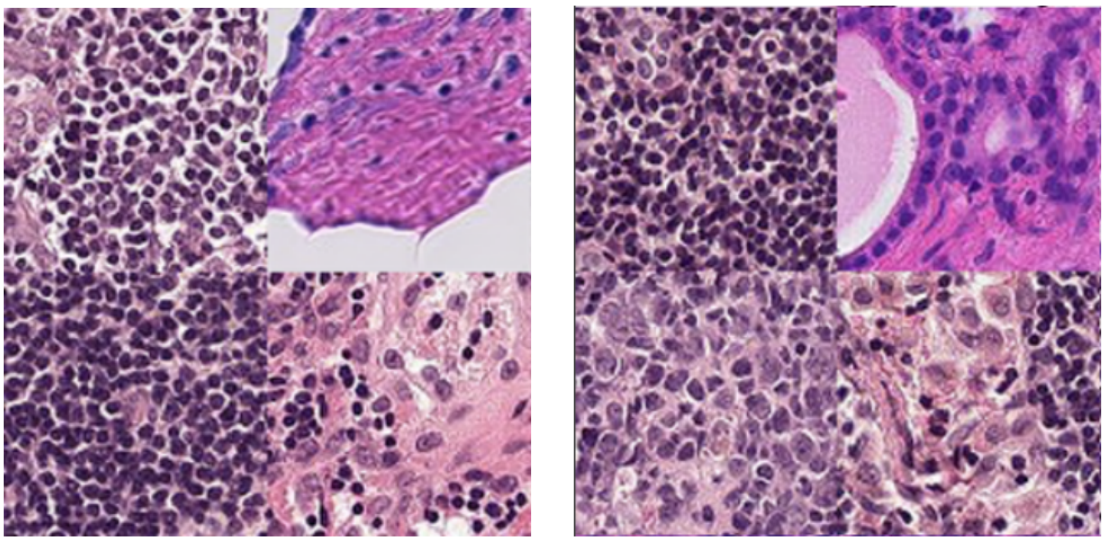
Paper Publicado: Histopathology Image Augmentation Through StyleGAN2-ADA

Los resultados obtenidos demostraron de manera consistente la efectividad de StyleGAN2-ADA para generar imágenes histopatológicas sintéticas de alta calidad y su impacto positivo en el rendimiento de los clasificadores. Los modelos generativos alcanzaron puntajes FID de 3,2 para el dataset PCam y 2,608 para IDC, indicando una alta fidelidad visual de las imágenes sintéticas generadas. La evaluación cualitativa realizada por un experto clínico confirmó que las imágenes generadas mantienen coherencia histopatológica y son visualmente similares a las muestras reales, validando su potencial utilidad en aplicaciones médicas.

No obstante, el experto también señaló una limitación importante: las resoluciones utilizadas son demasiado bajas para un uso clínico real. Al realizar acercamientos (zoom) sobre las imágenes, la pérdida de detalle es significativa, lo que imposibilita un análisis diagnóstico riguroso. En consecuencia, se vuelve imperativo trabajar con datasets de mayor resolución si se busca una integración o evaluación efectiva de estas técnicas en entornos médicos reales.

En el dataset PCam, los resultados mostraron mejoras significativas y consistentes en ambos es-

cenarios de disponibilidad de datos (Tabla 4.1). En el escenario más restrictivo con solo 3 % de datos disponibles, el modelo base alcanzó un AUC-ROC de 0,813, mientras que la implementación con aumento StyleGAN2-ADA logró 0.861, representando una mejora del 5,9 %. Esta mejora se vio acompañada de incrementos notables en las métricas específicas de la clase cáncer, con la exactitud aumentando de 0.86 a 0.89 y la sensibilidad mejorando sustancialmente de 0,7 a 0,81.



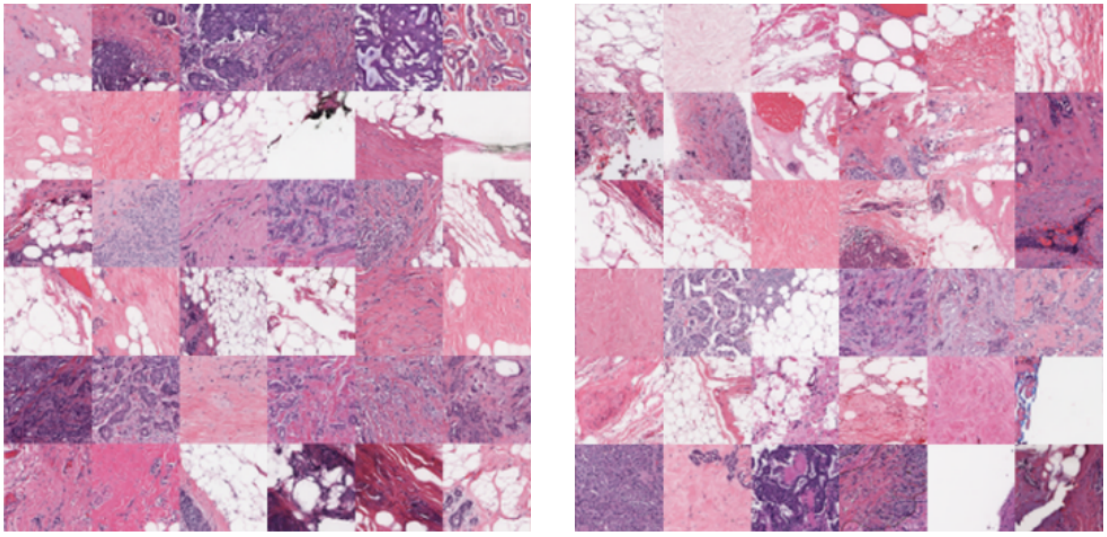
(a) Imágenes sintéticas (generadas) (b) Imágenes reales

Figura 4.1: Ejemplo de imágenes reales y sintéticas (generadas por el modelo) para el dataset PCAM. El puntaje FID obtenido es de 3,2.

Cuando se incrementó la disponibilidad de datos al 20 %, las mejoras fueron aún más pronunciadas, con el AUC-ROC aumentando de 0,797 a 0,887, una mejora del 11,3 %. Estos resultados sugieren que el aumento de datos sintéticos es particularmente efectivo cuando se combina con cantidades moderadas de datos reales.

Los resultados del dataset IDC revelaron patrones similares pero con características específicas relacionadas con el desbalance de clases inherente (Tabla 4.2). En el escenario de 3 % de disponibilidad de datos, el modelo base obtuvo un AUC-ROC de 0,856, que se incrementó a 0,885 con el aumento StyleGAN2-ADA, representando una mejora del 3,4 %. Sin embargo, las

mejoras más destacables se observaron en las métricas específicas de la clase cáncer, donde la exactitud aumentó dramáticamente de 0,75 a 0,89, la sensibilidad de 0,63 a 0,85, y el F1-score de 0,69 a 0,87. Con 20 % de disponibilidad de datos, el AUC-ROC mejoró de 0,862 a 0,883, una mejora del 2,4 %, manteniendo incrementos sustanciales en las métricas de detección de cáncer.



(a) Imágenes sintéticas (generadas) (b) Imágenes reales

Figura 4.2: Ejemplo de imágenes reales y sintéticas (generadas por el modelo) para el dataset IDC. El puntaje FID obtenido es de 2,603.

Un hallazgo particularmente relevante emergió del análisis del comportamiento del clasificador en el dataset IDC al 20 % de disponibilidad de datos. Mientras que las métricas de la clase cáncer mejoraron significativamente, se observó una reducción en las métricas de la clase no-cáncer. Este fenómeno, lejos de representar una degradación del rendimiento, indica una corrección del sesgo inicial del modelo hacia la clase mayoritaria. El incremento sostenido en el AUC-ROC confirma que el modelo mejorado es más discriminativo y equilibrado en su capacidad de clasificación, un aspecto crucial en aplicaciones médicas donde la detección de casos positivos (cáncer) tiene mayor importancia clínica que mantener alta precisión en casos negativos.

Los hallazgos clave del estudio establecieron que StyleGAN2-ADA demuestra particular efectividad en escenarios de datos extremadamente limitados, donde las mejoras en rendimiento

son más pronunciadas, como puede notarse en las Tablas 4.1 y 4.2. La mejora consistente en las métricas de detección de cáncer en ambos datasets valida la relevancia clínica del enfoque, especialmente considerando que en aplicaciones médicas la sensibilidad para detectar casos positivos es frecuentemente más crítica que otras métricas. Para más detalles, visitar el paper completo publicado en Springer [Muñoz et al., 2025b].

PCam	Modelo	ROC AUC	Clase: No-cancer			Clase: Cancer		
			Acc	Recall	F1	Acc	Recall	F1
3 %	Baseline	0.813	0.75	0.88	0.81	0.86	0.70	0.77
	+ StyleGAN2-ADA aumento, r = 0.3 + balance	**0.861**	**0.86**	**0.89**	**0.87**	**0.89**	**0.81**	**0.85**
20 %	Baseline	0.797	0.77	0.83	0.80	0.82	0.75	0.78
	+ StyleGAN2-ADA aumento, r = 0.3 + balance	**0.887**	**0.78**	**0.85**	**0.81**	**0.84**	**0.76**	**0.80**

Tabla 4.1: Resultados de los puntajes de clasificación con aumento generativo de datos usando StyleGAN2-ADA, con diferentes porcentajes de datos del dataset PCAM, separados por clase (cancer/no-cancer).

IDC	Modelo	ROC AUC	Clase: No-cancer			Clase: Cancer		
			Acc	Recall	F1	Acc	Recall	F1
3 %	Baseline	0.856	**0.89**	**0.93**	**0.91**	0.75	0.63	0.69
	+ StyleGAN2-ADA aumento, r = 0.3 + balance	**0.885**	0.85	0.89	0.87	**0.89**	**0.85**	**0.87**
20 %	Baseline	0.862	**0.87**	**0.96**	**0.91**	0.81	0.56	0.66
	+ StyleGAN2-ADA aumento, r = 0.3 + balance	**0.883**	0.84	0.90	0.87	**0.88**	**0.84**	**0.86**

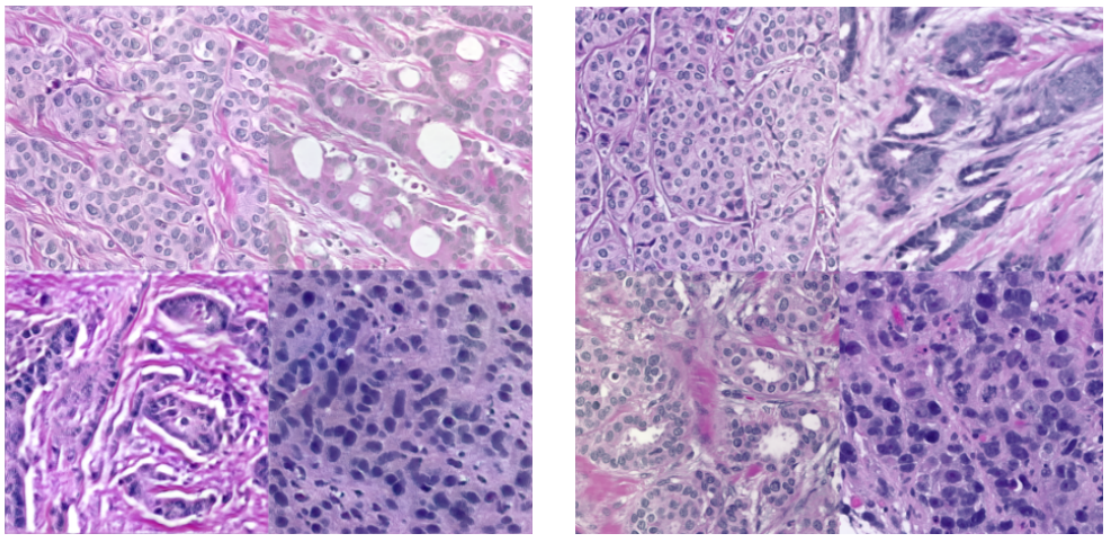
Tabla 4.2: Resultados de los puntajes de clasificación con aumento generativo de datos usando StyleGAN2-ADA, con diferentes porcentajes de datos del dataset IDC, separados por clase (cancer/no-cancer).

Pruebas con BreCaHAD y Análisis de Recursos Computacionales

Para el dataset BreCaHAD, se obtuvo un puntaje FID de 69.532, lo cual resulta llamativo, ya que indica una alta discrepancia entre las imágenes reales y las generadas, a pesar de que visualmente estas últimas no parecen tan diferentes. Esta aparente contradicción puede atribuirse a la alta resolución de las imágenes (512×512 píxeles), que introduce una gran cantidad de detalles sutiles difíciles de percibir visualmente, pero que el FID sí logra capturar. Estos detalles imperceptibles reflejan deficiencias en la calidad de las imágenes sintéticas, posiblemente causadas por una cantidad limitada de datos de entrenamiento y tiempo insuficiente para que el modelo converja adecuadamente.

Adicionalmente, la gran resolución de las imágenes de BreCaHAD impactó significativamente en el tiempo de entrenamiento: el modelo requirió aproximadamente 2 días, 22 horas y 21

minutos para procesar 1.000 kimg (donde 1 kimg equivale a mil imágenes). En contraste, el dataset PCam procesó 25.000 kimg en 3 días, 9 horas y 6 minutos, mientras que IDC alcanzó la misma cantidad en tan solo 15 horas y 9 minutos. Estos resultados evidencian la complejidad computacional exponencial que implica duplicar la resolución de las imágenes al entrenar modelos generativos. De hecho, ya existen estudios que cuantifican estos costos en función de la resolución, el tipo y la cantidad de GPUs empleadas, como se muestra en la Tabla B.1.



(a) Imágenes sintéticas (generadas) (b) Imágenes reales

Figura 4.3: Ejemplo de imágenes reales y sintéticas (generadas por el modelo) para el dataset BreCaHAD. El puntaje FID obtenido es de 69,532.

El análisis de los recursos computacionales requeridos reveló la importancia de infraestructuras de computación de alto rendimiento para proyectos de esta naturaleza. Además, este análisis permitió establecer una cota superior respecto a la posible resolución de imagenes a utilizar, pues una resolución de 512×512 o más, junto con los recursos computacionales disponibles iposibilita el entrenamiento del modelo generativo. Esto, junto con el feedback propuesto por el experto, hacen que sea necesario un dataset con una resolución mínima tal que se pueda efectuar un diagnóstico. En este contexto, se optó por el dataset *NCT-CRC-HE-100k*, que cumple con estos requisitos.

La variación de los tiempos de entrenamiento se puede ver en la Tabla 4.3, utilizando una sola GPU NVIDIA Tesla V100.

Pruebas de Manipulación del Espacio Latente

Se realizó un mini-experimento para mezclar estilos entre pares de imágenes generadas por la GAN. Este consiste en mezclar los códigos latentes de cada imagen en diferentes escalas para conocer qué partes de la imagen se modifican al perturbar un determinado subconjunto de estilos. Se generan dos conjuntos de imágenes desde sus respectivos códigos latentes, fuente A y fuente B, que hacen referencia a la fila 0 y columna 0, respectivamente. El resto de las imágenes son generadas copiando un subconjunto de estilos específicos desde B y tomando el resto desde A.

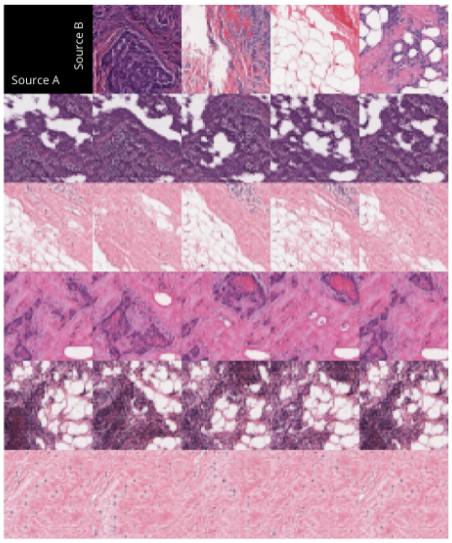
La Figura 4.4a muestra que copiar estilos correspondientes a las resoluciones bajas o gruesas $4^2 – 8^2$ desde B, cambia la forma de las estructuras de manera muy leve y se mantiene el color y aspectos de alto nivel, que podríamos decir que sigue siendo el mismo tipo o parte del tejido de A.

Por otro lado, la Figura 4.4b muestra que copiar estilos correspondientes a las resoluciones intermedias $16^2 – 64^2$ desde B, cambia en mayor proporción las estructuras, manteniendo una predominancia por aquellas presentes en B, pero aún manteniendo el color y tipo o parte del tejido de A.

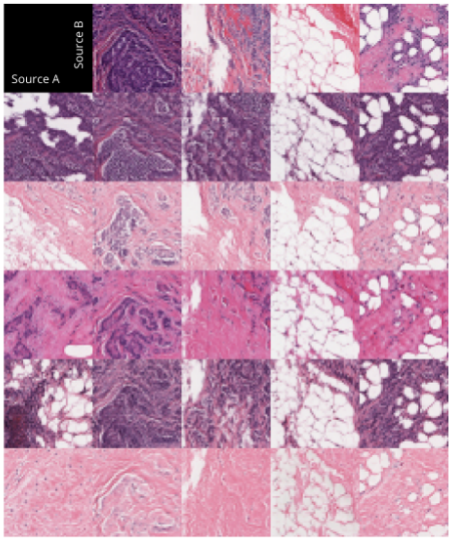
Finalmente, la Figura 4.5a muestra que copiar estilos correspondientes a las resoluciones finas $64^2 – 256^2$ desde B, cambia el color y tipo de tejido o parte del tejido, manteniendo una predominancia por aquellas presentes en B, pero aún manteniendo la forma de las estructuras de la imagen A.

Estos resultados establecieron una base sólida para el desarrollo posterior de métodos de explicabilidad, demostrando que StyleGAN2-ADA no solo es capaz de generar imágenes histopatológicas realistas, sino que también posee la capacidad de manipular efectivamente el espacio latente para producir variaciones controladas. Esta capacidad de manipulación del espacio latente representa el nexo fundamental con el trabajo de explicabilidad contrafactual, donde la generación de

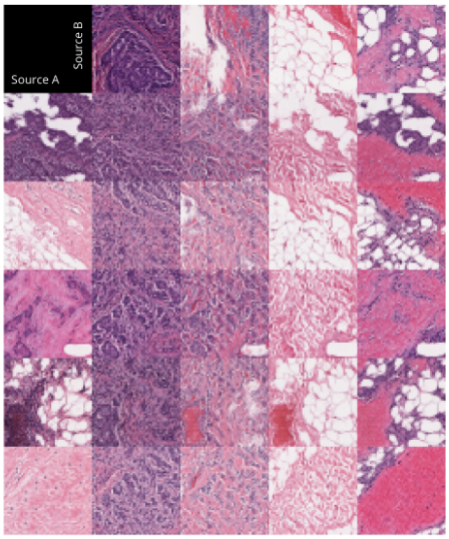
contrafactuales requiere modificaciones precisas y controladas de las representaciones latentes para producir explicaciones visuales interpretables.



(a) Copia de estilos correspondientes a las resoluciones gruesas o bajas de la fuente B en A.



(b) Copia de estilos correspondientes a las resoluciones intermedias de la fuente B en A.



(a) Copia de estilos correspondientes a las resoluciones finas de la fuente B en A.

Dataset	Resolución	# imágenes	FID	Tiempo de entrenamiento
BreCaHad (full-set)	512×512	162 (100 %)	69,532	2d 22h 21m
NCT-CRC-HE (subset)	256×256	17526 (17,5 %)	16,210	5d 12h 38m
PCam (subset)	128×128	52429 (20 %)	3,200	3d 09h 06m
PCam (subset)	128×128	7864 (3 %)	3,738	1d 02h 24m
IDC (subset)	64×64	37789 (20 %)	2,608	1d 14h 00m
IDC (subset)	64×64	5668 (3 %)	2,603	0d 15h 09m

Tabla 4.3: Configuración y resultados de entrenamiento del modelo StyleGAN2-ADA con 4 datasets diferentes; PCam, IDC, BreCaHAD y NCT-CRC-HE. El entrenamiento fue ejecutado en una sola GPU V100 del Laboratorio Nacional para Computación de Alto Rendimiento (NLHPC).

4.2. Resultados del método generativo de imágenes

Resultados y Estrategia de Entrenamiento

En esta sección, presentamos los resultados obtenidos al entrenar el modelo propuesto, compuesto por un generador StyleGAN2-ADA, un codificador (encoder) y un clasificador, utilizando el conjunto de datos *NCT-CRC-HE*. El objetivo principal era lograr una alta calidad de generación de imágenes, con un espacio latente desentrelazado y guiado por el clasificador, al mismo tiempo que se obtienen reconstrucciones de imágenes de alta calidad mediante el encoder. La función de pérdida total utilizada es la ecuación 3.6 descrita en la Sección 3:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}},$$

donde la pérdida de reconstrucción se descompone en:

$$\mathcal{L}_{\text{rec}} = \lambda_x \mathcal{L}_{\text{rec\_x}} + \lambda_w \mathcal{L}_{\text{rec\_w}} + \lambda_{\text{lpips}} \mathcal{L}_{\text{lpips}}.$$

Estrategia de Entrenamiento Optimizada

Inicialmente, se realizaron pruebas ponderando equitativamente las pérdidas de clasificación y reconstrucción (i.e., $\lambda_{\text{cls}} = \lambda_{\text{rec}} = 1$) desde el inicio del entrenamiento (ver Tabla 4.4). Sin embargo, se observó que esta estrategia comprometía la calidad final de la reconstrucción, ya que el gradiente de clasificación influía significativamente desde las primeras etapas.

Se descubrió una estrategia más efectiva consistente en un enfoque de dos etapas (ver Tabla 4.5):

1. **Pre-entrenamiento de Reconstrucción:** Entrenar el encoder y el generador (si se entrena conjuntamente) utilizando únicamente la pérdida de reconstrucción $\mathcal{L}_{\text{rec}}$ (o con un λ_{cls} muy cercano a cero). Esto permite al encoder aprender a invertir las imágenes en el espacio latente de manera eficaz sin la interferencia inicial de la tarea de clasificación.
2. **Entrenamiento Conjunto:** Una vez alcanzada una calidad de reconstrucción base satisfactoria, se introduce o incrementa λ_{cls} a su valor objetivo (e.g., $\lambda_{\text{cls}} = 1$, manteniendo $\lambda_{\text{rec}} = 1$) y se continúa el entrenamiento con la pérdida $\mathcal{L}_{\text{total}}$ completa.

Este enfoque de dos etapas resultó en una mejora notable de las métricas de reconstrucción (especialmente LPIPS, PSNR y SSIM) en comparación con el entrenamiento conjunto desde el inicio. Además, produce una reducción significativa en la métrica FID independiente de la configuración de reconstrucción, lo que se traduce en una mejora en la calidad de las imágenes generadas por el método. Todos los experimentos detallados a continuación utilizaron esta estrategia optimizada.

Tabla 4.4: Comparación cuantitativa de métricas clave para diferentes configuraciones de pesos en la pérdida de reconstrucción ($\mathcal{L}_{rec}$). Todas las configuraciones utilizaron la estrategia de entrenamiento no optimizada en una sola etapa con λ_{cls} = 1 y λ_{rec} = 1 en toda la fase de entrenamiento. ↓ indica que valores más bajos son mejores, ↑ indica que valores más altos son mejores.

Config.	**Pesos Internos**			**Métricas Reconstrucción**			**Pérdida**	**Calidad**
	$\boldsymbol{\lambda_x}$	$\boldsymbol{\lambda_w}$	$\boldsymbol{\lambda_{lpips}}$	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	$\boldsymbol{\mathcal{L}_{rec\_w}}$ ↓	FID ↓
Baseline	1	1	10	0,25	40	0,9	Bajo (0,12)	**24,4**
Pixel-Dominant	10	0,1	0,1	0,3	**42,7**	**1,1**	Alto (0,41)	55,8
Perceptual	0,1	0,1	21,3	**0,22**	35,1	0,83	Moderado (0,19)	25,9
Latent	1	10	5	0,29	38,3	0,88	**Muy Bajo (0,07)**	27,5

Tabla 4.5: Comparación cuantitativa de métricas clave para diferentes configuraciones de pesos en la pérdida de reconstrucción ($\mathcal{L}_{rec}$). Todas las configuraciones utilizaron la estrategia de entrenamiento optimizada en dos etapas con λ_{cls} = 1 y λ_{rec} = 1 en la fase final. ↓ indica que valores más bajos son mejores, ↑ indica que valores más altos son mejores.

Config.	**Pesos Internos**			**Métricas Reconstrucción**			**Pérdida**	**Calidad**
	$\boldsymbol{\lambda_x}$	$\boldsymbol{\lambda_w}$	$\boldsymbol{\lambda_{lpips}}$	LPIPS ↓	PSNR (dB) ↑	SSIM ↑	$\boldsymbol{\mathcal{L}_{rec\_w}}$ ↓	FID ↓
Baseline	1	1	10	0,15	28,0	0,80	Bajo (0,05)	**16,2**
Pixel-Dominant	10	0,1	0,1	0,20	**29,5**	**0,82**	Alto (0,15)	35
Perceptual	0,1	0,1	20	**0,10**	25	0,75	Moderado (0,10)	18,5
Latent	1	10	5	0,18	27	0,78	**Muy Bajo (0,02)**	19,9

Comparación de Configuraciones de Pérdida de Reconstrucción

Para investigar el impacto de los componentes individuales de $\mathcal{L}_{\text{rec}}$, se evaluaron distintas configuraciones de los pesos internos λ_x, λ_w, λ_{lpips}, manteniendo $\lambda_{\text{rec}} = 1{,}0$ y $\lambda_{\text{cls}} = 1{,}0$ (en la segunda etapa del entrenamiento). A continuación, se describen los resultados de tres configuraciones representativas:

Configuración 1: Baseline Equilibrado ($\lambda_x = 1$ | $\lambda_w = 1$ | $\lambda_{\textbf{lpips}} = 10$)

Esta configuración buscaba un balance entre las tres componentes. Las reconstrucciones obtenidas mostraron una buena calidad perceptual general, como lo indica un valor bajo de LPIPS (e.g., ~0,15). La fidelidad a nivel de píxeles fue razonable, con valores moderados de PSNR (e.g., ~28 dB) y SSIM (e.g., ~0,80). La pérdida de reconstrucción latente $\mathcal{L}_{\text{rec\_w}}$ también convergió a valores bajos, sugiriendo una inversión consistente en el espacio W. Visualmente (ver Figura 4.6), las imágenes son fieles a las originales, aunque pueden presentar una ligera suavización en texturas muy finas.

Configuración 2: Dominancia de Pixel ($\lambda_x = 10$ | $\lambda_w = 0,1$ | $\lambda_{\textbf{lpips}} = 1$)

Esta configuración enfatiza fuertemente la coincidencia a nivel de píxeles al asignar un peso dominante a $\mathcal{L}_{\text{rec\_x}}$, mientras que las pérdidas latente y perceptual tienen una influencia mínima. Como era de esperarse, el entrenamiento fue dominado por el gradiente de la pérdida L1, con el optimizador enfocándose casi exclusivamente en minimizar las diferencias de intensidad entre la imagen original y la reconstruida.

Las reconstrucciones resultantes presentaron una fidelidad numérica muy alta: se obtuvieron valores elevados de PSNR (e.g., ~30 dB) y SSIM (e.g., ~0,85), confirmando una coincidencia precisa a nivel de píxeles. No obstante, esta precisión no se tradujo en una mejor calidad perceptual. El valor de LPIPS fue considerablemente más alto (e.g., ~0,22), lo que indica una percepción visual más pobre. Este fenómeno también se reflejó visualmente (*ver Figura 4.7*), donde las imágenes reconstruidas aparecen borrosas, especialmente en regiones con texturas

finas o patrones de alto detalle.

Este resultado es coherente con hallazgos previos en la literatura sobre optimización en el espacio de píxeles usando L1/L2, donde las reconstrucciones tienden a promediar múltiples modos posibles, resultando en imágenes perceptualmente planas o sin vida.

Adicionalmente, la pérdida de reconstrucción latente $\mathcal{L}_{\text{rec\_w}}$ fue más alta que en otras configuraciones, lo que sugiere que el encoder genera vectores latentes menos consistentes o más ruidosos.

En conjunto, aunque la configuración de dominancia pixel logra una gran coincidencia numérica, sacrifica tanto la consistencia latente como la fidelidad perceptual, haciendo de esta configuración una opción subóptima para tareas donde la calidad visual o la interpretabilidad de las manipulaciones latentes son importantes.

Configuración 3: Dominancia Latente ($\lambda_x = 1$ | $\lambda_w = 10$ | $\lambda_{\textbf{lpips}} = 5$)

El objetivo en este caso fue lograr que el proceso de inversión hacia el espacio latente W fuera más consistente, es decir, que las imágenes editadas se mantuvieran lo más fieles posible a las originales. Para ello, se incrementó de forma considerable el valor de λ_w. Como era de esperar, esta configuración logró la menor pérdida de reconstrucción latente $\mathcal{L}_{\text{rec\_w}}$. Esto sugiere que el encoder transforma las imágenes a representaciones latentes más estables o predecibles, lo cual podría ser beneficioso para tareas de edición posteriores. Sin embargo, esta fuerte restricción en el espacio W pareció limitar ligeramente la capacidad del modelo para optimizar la similitud visual directa. Las métricas LPIPS (e.g., ~0,18) y PSNR (e.g., ~27 dB) fueron ligeramente peores que en la configuración baseline, indicando un compromiso entre la consistencia latente y la fidelidad de la reconstrucción visual (*ver Figura 4.8*).

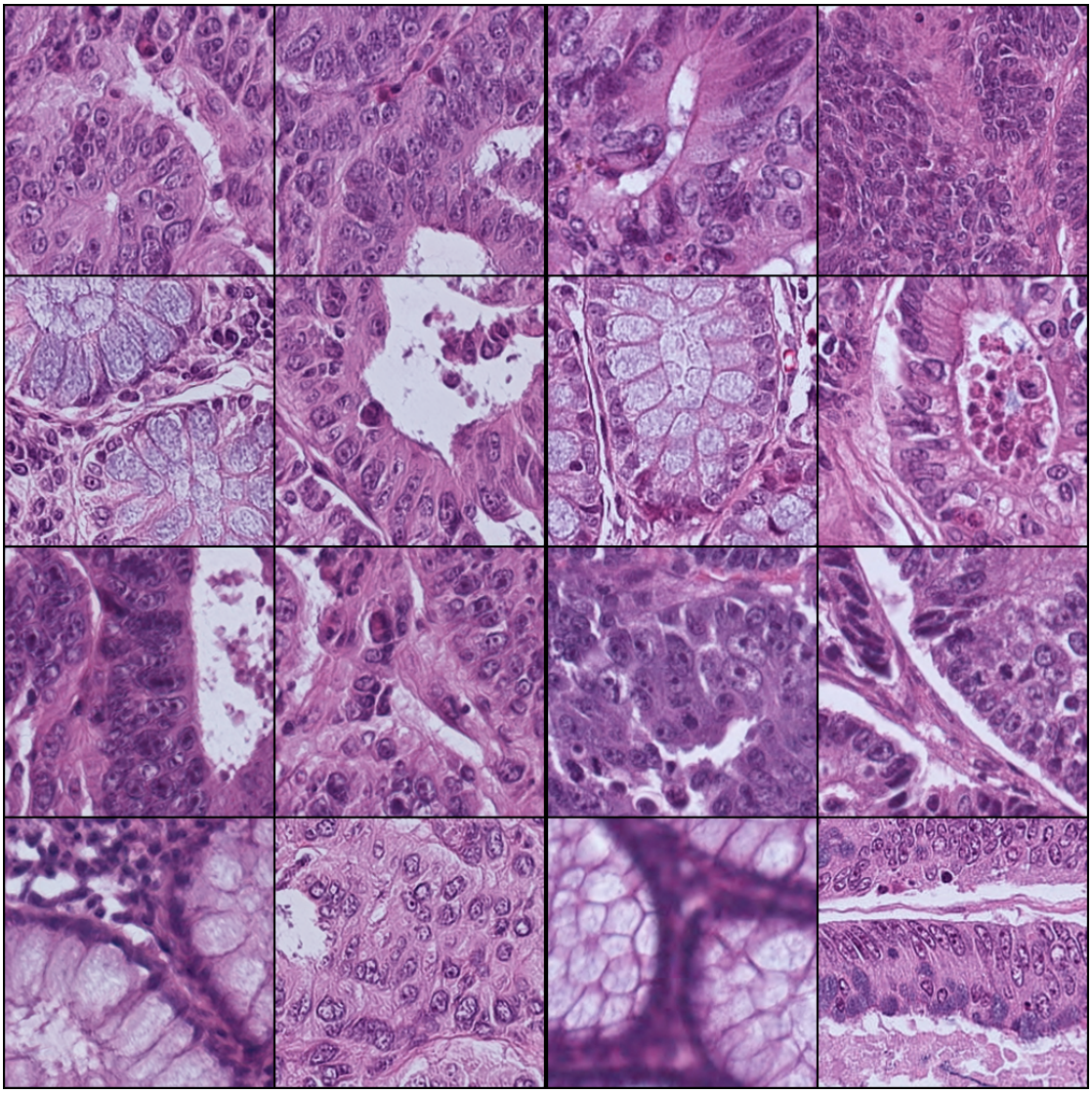


Figura 4.6: **Imágenes de reconstrucciones generadas por el método para la configuración 1: Baseline Equilibrado.** Las dos primeras columnas representan las reconstrucciones de las siguientes dos, respectivamente.

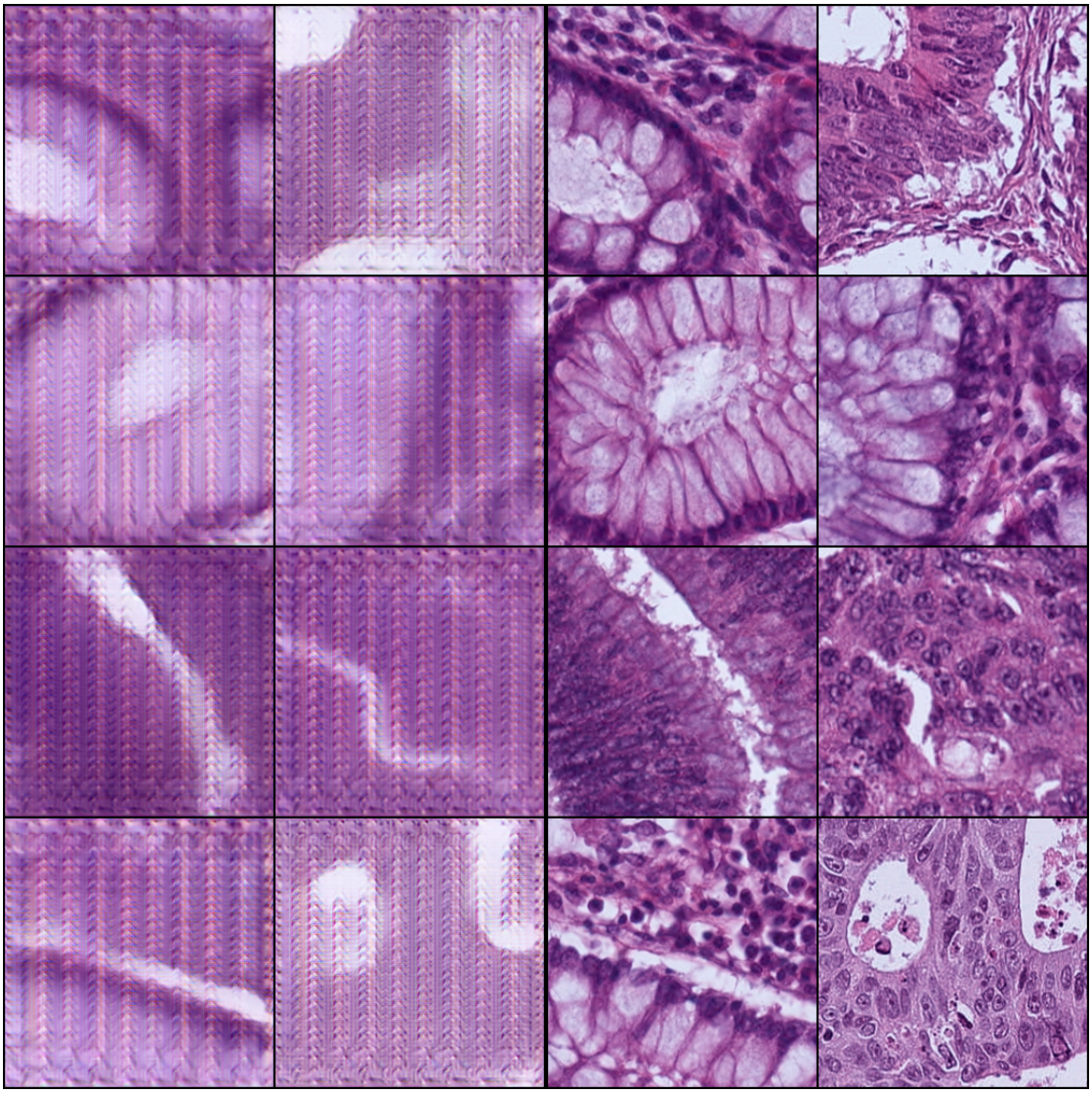


Figura 4.7: **Imágenes de reconstrucciones generadas por el método para la configuración 2: Dominancia de pixel.** Las dos primeras columnas representan las reconstrucciones de las siguientes dos, respectivamente.

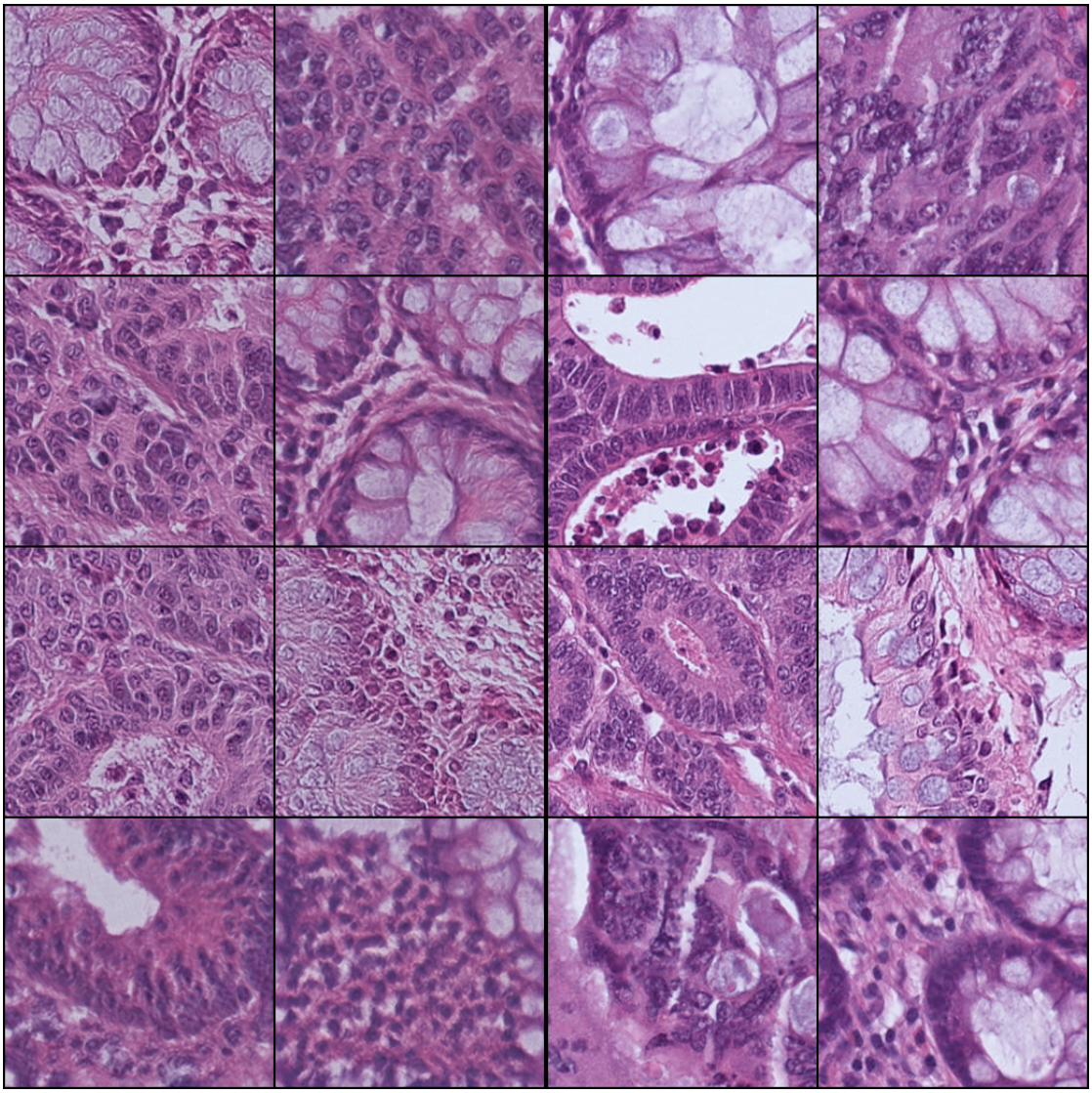


Figura 4.8: **Imágenes de reconstrucciones generadas por el método para la configuración 3: Dominancia Latente.** Las dos primeras columnas representan las reconstrucciones de las siguientes dos, respectivamente.

Discusión

Los resultados indican que la elección de los pesos internos de la pérdida de reconstrucción implica un claro *trade-off* entre fidelidad numérica, calidad perceptual y consistencia latente. La configuración baseline ofrece un buen compromiso general entre estos aspectos, sirviendo como punto de equilibrio para tareas generales de reconstrucción.

Priorizar la consistencia latente (Configuración 3) mejora la estabilidad de la inversión en el espacio W, lo cual es deseable para tareas de edición o generación controlada, pero puede degradar levemente la apariencia visual de las reconstrucciones.

Por otro lado, la configuración dominada por la pérdida de píxeles (Configuración 2) logra la mayor fidelidad numérica, con excelentes métricas PSNR y SSIM. Sin embargo, estas imágenes tienden a ser perceptualmente planas o borrosas, especialmente en regiones de alta frecuencia, y muestran menor consistencia latente. Esto refleja una limitación conocida de las pérdidas puramente pixel-wise cuando se aplican en modelos generativos.

La elección óptima depende del objetivo final del sistema. Para este proyecto, se adoptó la configuración 1 en los experimentos posteriores, ya que proporciona un rendimiento equilibrado entre fidelidad visual, precisión estructural y capacidad de edición latente. En particular, se observó que una estrategia de entrenamiento en dos etapas mejora de forma sustantiva la calidad de la reconstrucción independientemente de la configuración específica de pesos, al permitir que el encoder aprenda primero una inversión más robusta antes de abordar la clasificación.

4.3. Resultados de la Exploración del Espacio Latente

En esta sección, se presentan los resultados obtenidos del análisis y la exploración de los espacios latentes de la red generativa entrenada. El objetivo principal de esta exploración es comprender cómo el modelo organiza internamente las características aprendidas de las imágenes histopatológicas y si esta organización se correlaciona con atributos visuales semánticamente relevantes para la distinción entre tejido benigno y canceroso. La incorporación de una red ge-

nerativa basada en StyleGAN2-ADA en la arquitectura busca no solo generar imágenes realistas, sino también aprender una representación desentrelazada de las características tisulares.

Selección y Caracterización Inicial del Espacio Latente W

La arquitectura StyleGAN, introduce una distinción fundamental entre el espacio latente inicial Z y el espacio latente intermedio W. El espacio Z es típicamente un vector muestreado de una distribución simple (e.g., Gaussiana), mientras que el espacio W se obtiene a través de una red de transformación no lineal (*mapping network*) a partir de Z. Esta transformación está diseñada para desentrelazar los factores de variación presentes en los datos, lo que idealmente resulta en un espacio W donde las características semánticas son más linealmente separables y controlables.

Para evaluar la idoneidad de estos espacios para el análisis de interpretabilidad en el contexto histopatológico, se realizó una visualización comparativa utilizando la técnica de reducción de dimensionalidad Uniform Manifold Approximation and Projection (UMAP). La Figura 4.9 muestra las proyecciones UMAP de los espacios W (izquierda) y Z (derecha). En ambas visualizaciones, cada punto representa una imagen y su color indica la probabilidad de pertenencia a una clase según la salida de un clasificador entrenado sobre estas representaciones (1 para cancerosa - rojo, y 0 para benigna - azul).

- **Espacio W reducido con UMAP:** Se observa una clara separación entre las nubes de puntos correspondientes a la clase cancerosa (roja) y la clase benigna (azul). Aunque existe cierta superposición, la estructura global sugiere que el espacio W organiza las muestras de manera que las clases son considerablemente distinguibles. Esta separación es un indicativo del mayor poder de representación y desentrelazado del espacio W.
- **Espacio Z reducido con UMAP:** En contraste, la proyección del espacio Z muestra una mezcla mucho más pronunciada de puntos rojos y azules. La separación entre clases es significativamente menos evidente, lo que sugiere que los factores de variación relacionados con cada clase no están tan claramente desentrelazados en este espacio original.

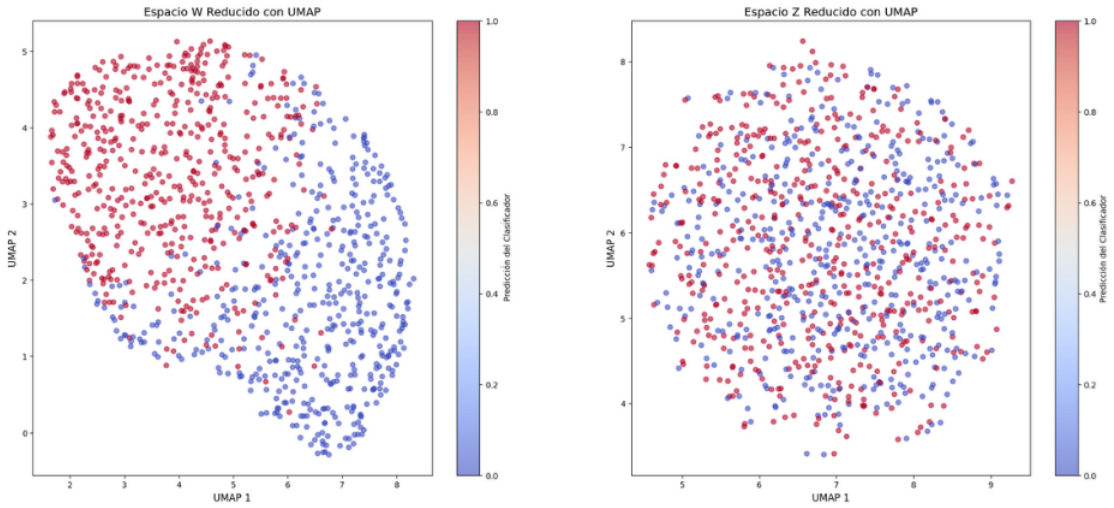


Figura 4.9: **Visualización de los espacios latentes de W y Z obtenidas mediante UMAP.** El color de cada punto indica la probabilidad de pertenencia a una clase según la salida del clasificador, 1 para cancerosa y 0 para benigna.

Estos resultados confirman empíricamente las ventajas teóricas del espacio W. La arquitectura propuesta, en particular, la red de transformación de StyleGAN2-ADA junto con el clasificador, han logrado transformar el espacio Z, más enredado, en un espacio W donde las características relevantes para la clasificación histopatológica están mejor estructuradas y son más accesibles para el análisis. Por consiguiente, la exploración detallada se centrará en el espacio latente W.

La Figura 4.10 proporciona una visualización adicional del espacio latente W, esta vez superponiendo una estimación de la densidad de los datos sobre la proyección UMAP. Los colores de los puntos siguen representando la probabilidad de pertenencia a la clase cancerosa (rojo) o benigna (azul). Esta visualización no solo confirma la separación de clases observada en la Figura 4.9, sino que también resalta las regiones de mayor concentración de datos dentro del espacio W (regiones de color amarillo). Se pueden apreciar contornos de densidad que delimitan las áreas donde el modelo ha transformado la mayoría de las muestras de entrenamiento. Las zonas de alta densidad sugieren la existencia de prototipos o características en común en el conjunto de datos, que el modelo ha aprendido a representar de forma compacta.

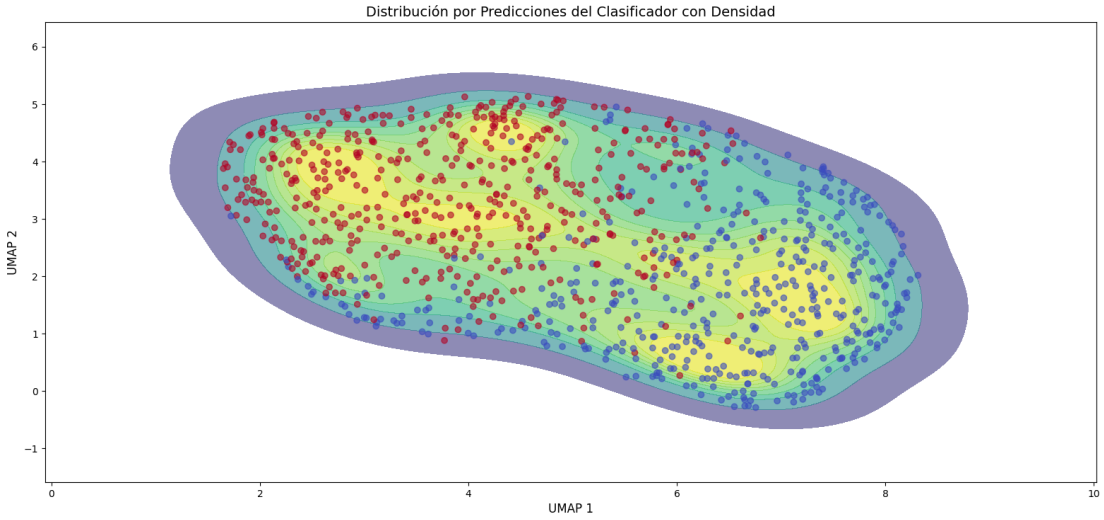


Figura 4.10: **Visualización del espacio latente de W junto con la densidad de los datos.**

La exploración subsiguiente se enfocará en analizar con mayor detalle regiones específicas dentro de esta distribución de densidad del espacio W, examinando las características histopatológicas de las imágenes correspondientes a clústeres identificados dentro de cada clase.

Análisis del Espacio Latente W para la Clase Benigna

La Figura 4.11 muestra la proyección UMAP del espacio latente W para las imágenes pertenecientes a la clase benigna. En esta proyección, se identificaron varias regiones de interés, usualmente las regiones de mayor densidad, que se indican con puntos resaltados en azul y etiquetadas como clústeres A, B, C y D. A la derecha de la proyección UMAP, se presentan ejemplos de imágenes histopatológicas correspondientes a las imágenes que componen cada uno de estos clústeres. En la Figura 4.12 se muestra la misma la misma proyección pero con una leyenda que indica el número de punto junto con la etiqueta predicha por el clasificador.

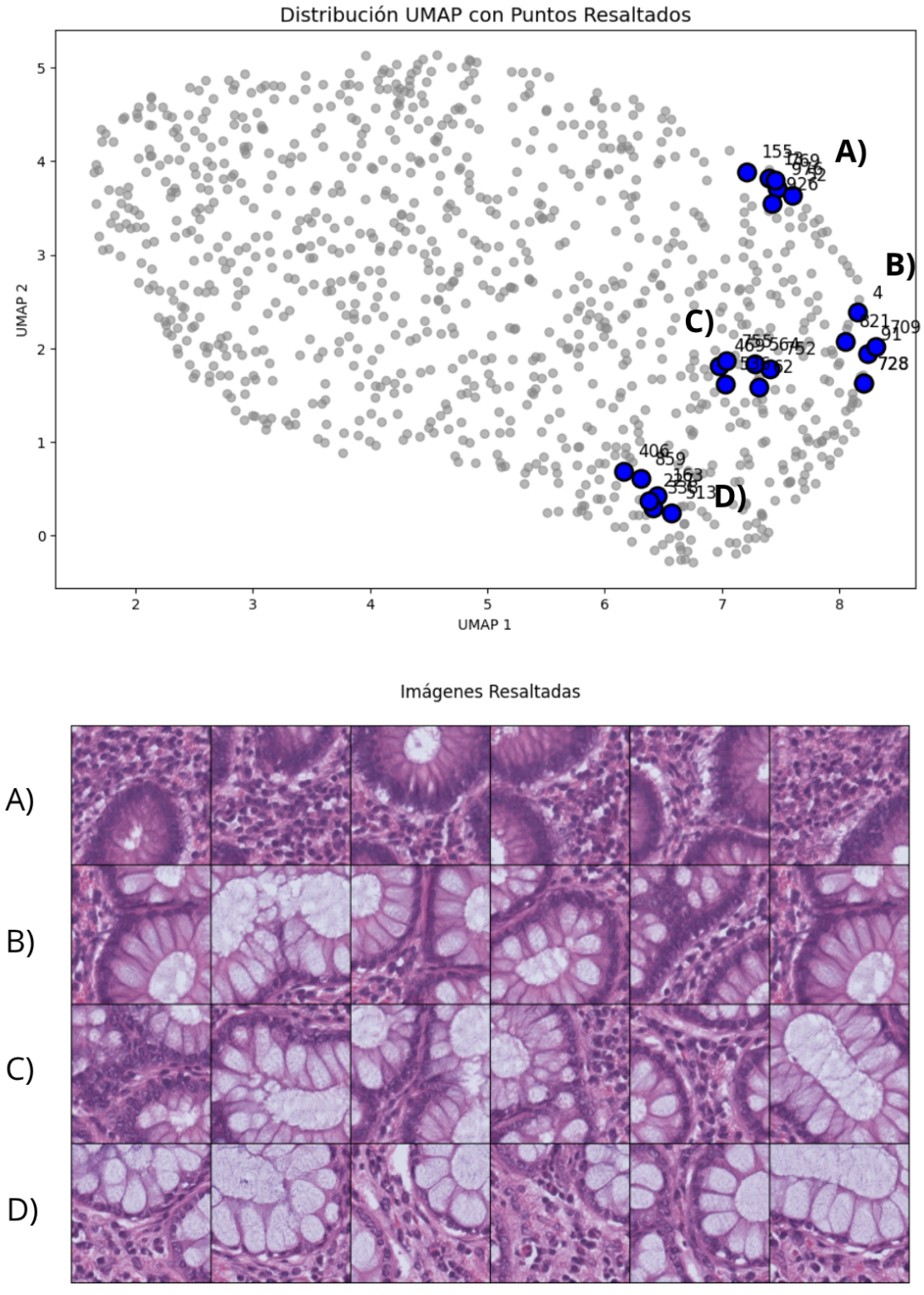


Figura 4.11: **Visualización del espacio latente de W mediante UMAP, donde se han resaltado grupos de puntos correspondientes a la clase benigna (0).** Las imágenes correspondientes a estos puntos (A, B, C, D) se muestran a la derecha, ilustrando el tipo de características representadas en esas regiones del espacio latente.

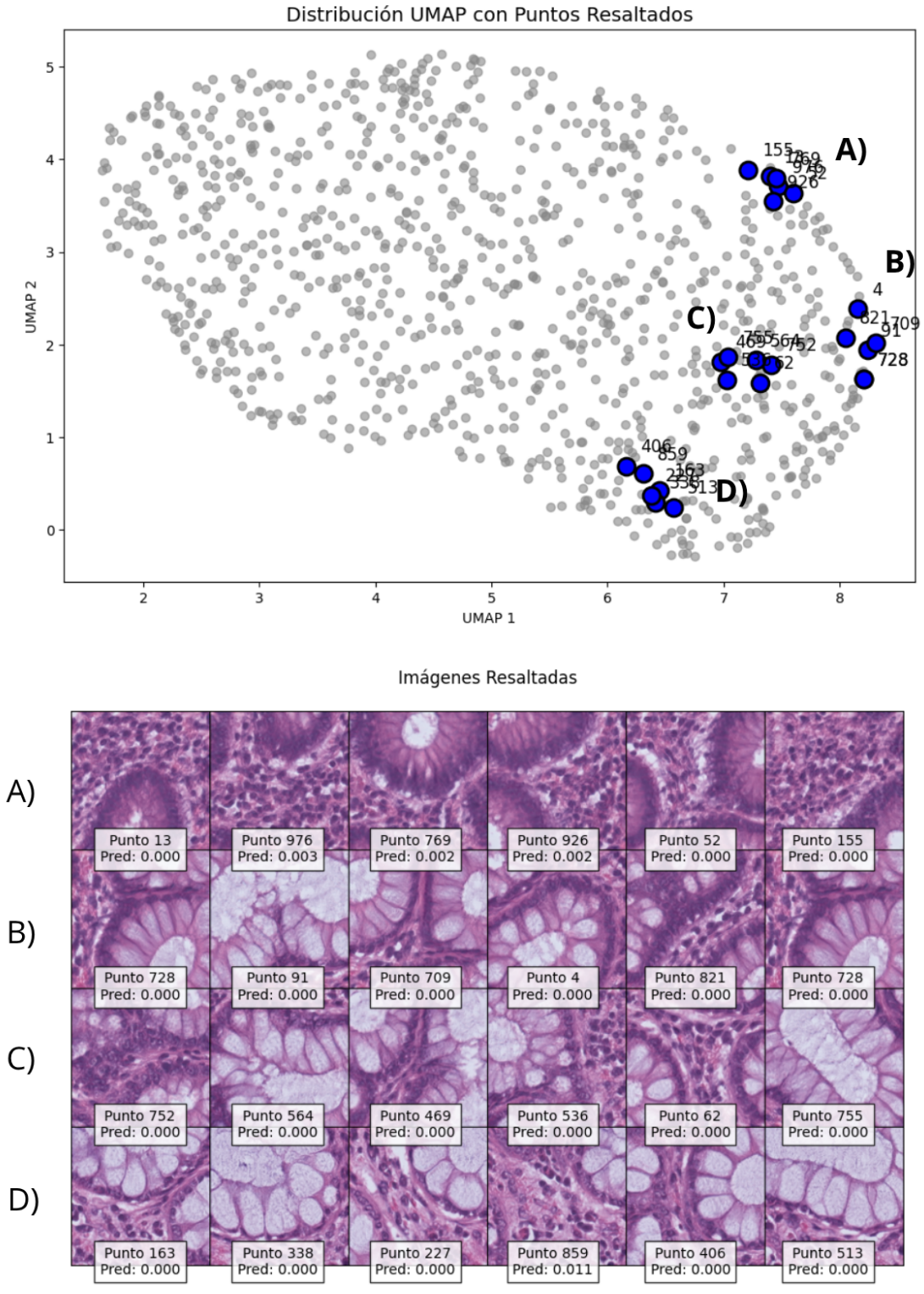


Figura 4.12: **Visualización del espacio latente de W mediante UMAP para la clase benigna, pero con el detalle de cada imagen.** Misma visualización del espacio latente de W mediante UMAP para la clase benigna, pero con el detalle de cada imagen (predicción y número de cada punto proyectado.)

Una observación general es que, efectivamente, **las imágenes agrupadas dentro de un mismo clúster comparten características morfológicas distintivas y comunes entre sí**. Por ejemplo, todas las imágenes son consistentes con tejido benigno, muestran predominantemente estructuras glandulares bien formadas, con una celularidad moderada y arquitectura tisular preservada. Además, se observa una consistencia en el tamaño y forma de las glándulas a nivel general. No obstante, y pese a que estamos en un espacio latente que corresponde a una única clase, hay diferencias graduales entre los distintos clústers; desde A hacia D podemos ver una apertura gradual de las glándulas o estructuras más elongadas, un mayor aumento de túbulos y una tendencia hacia un color cada vez más claro.

Es importante destacar que, si bien cada clúster codifica un conjunto de características predominantes, existe una transición gradual y variaciones sutiles al moverse dentro de un mismo clúster y entre clústeres adyacentes. **Esto sugiere que el espacio latente W ha aprendido a representar un espectro continuo de las variaciones morfológicas presentes en el tejido benigno**. La diferenciación en subgrupos (A, B, C, D) indica una codificación de características específicas y discernibles, lo que demuestra la capacidad del modelo para capturar la heterogeneidad inherente a esta clase. Cada letra, representando una fila de imágenes, subraya cómo el modelo agrupa patrones visuales con un alto grado de similitud, por ejemplo, la forma y el espaciado de las glándulas, o la textura del tejido circundante.

Análisis del Espacio Latente W para la Clase Cancerosa

De manera análoga, se realizó el análisis para la clase cancerosa, cuyos resultados se visualizan en la Figura 4.13. Aquí, los puntos resaltados en rojo identifican regiones de alta densidad en la proyección UMAP, etiquetadas como clústeres A, B y C. Las imágenes correspondientes a estos clústeres se muestran a la derecha de cada figura.

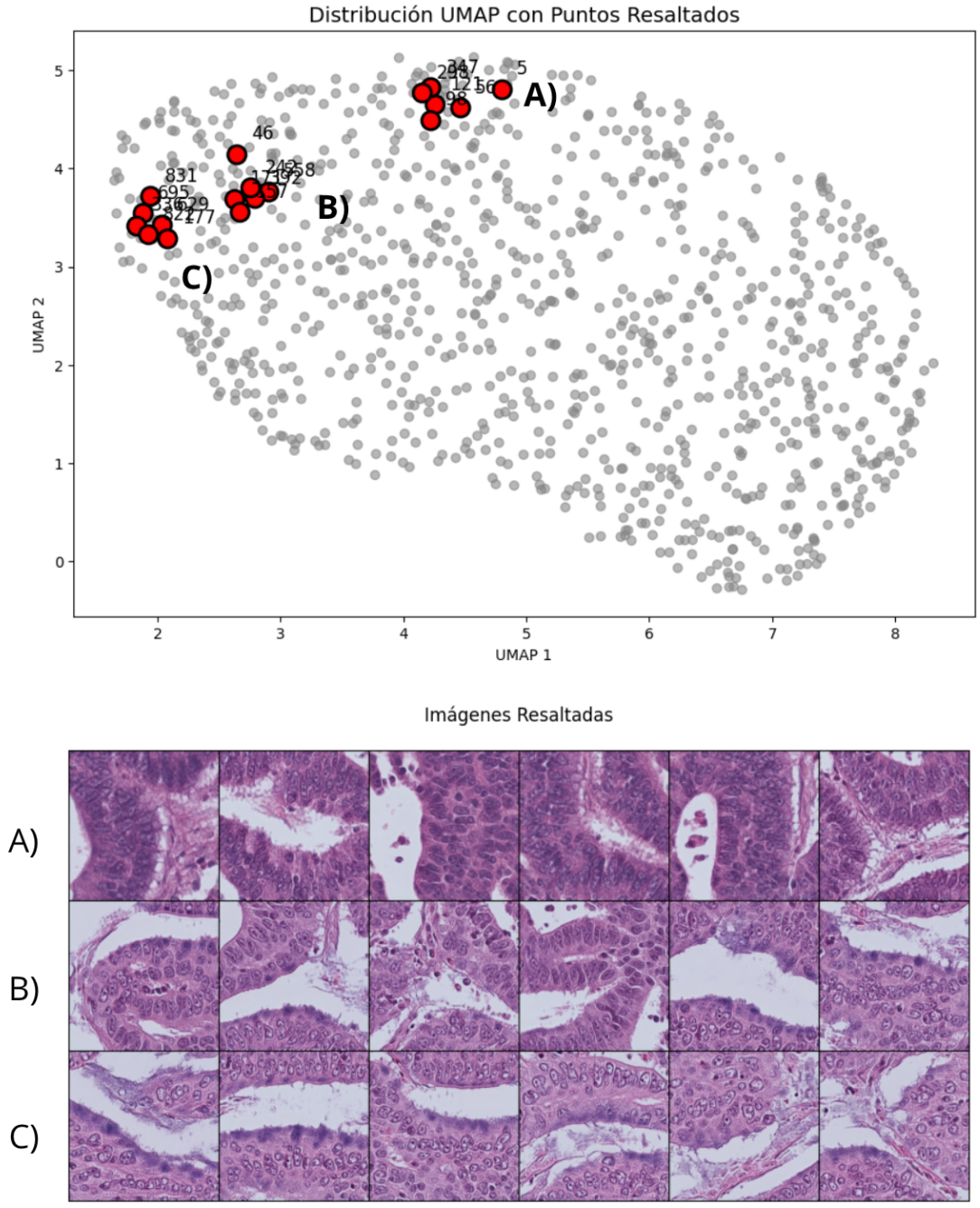


Figura 4.13: **Visualización del espacio latente de W mediante UMAP, donde se han resaltado grupos de puntos correspondientes a la clase cancerosa (1).** Las imágenes correspondientes a estos puntos (A, B, C) se muestran a la derecha, ilustrando el tipo de características representadas en esas regiones del espacio latente.

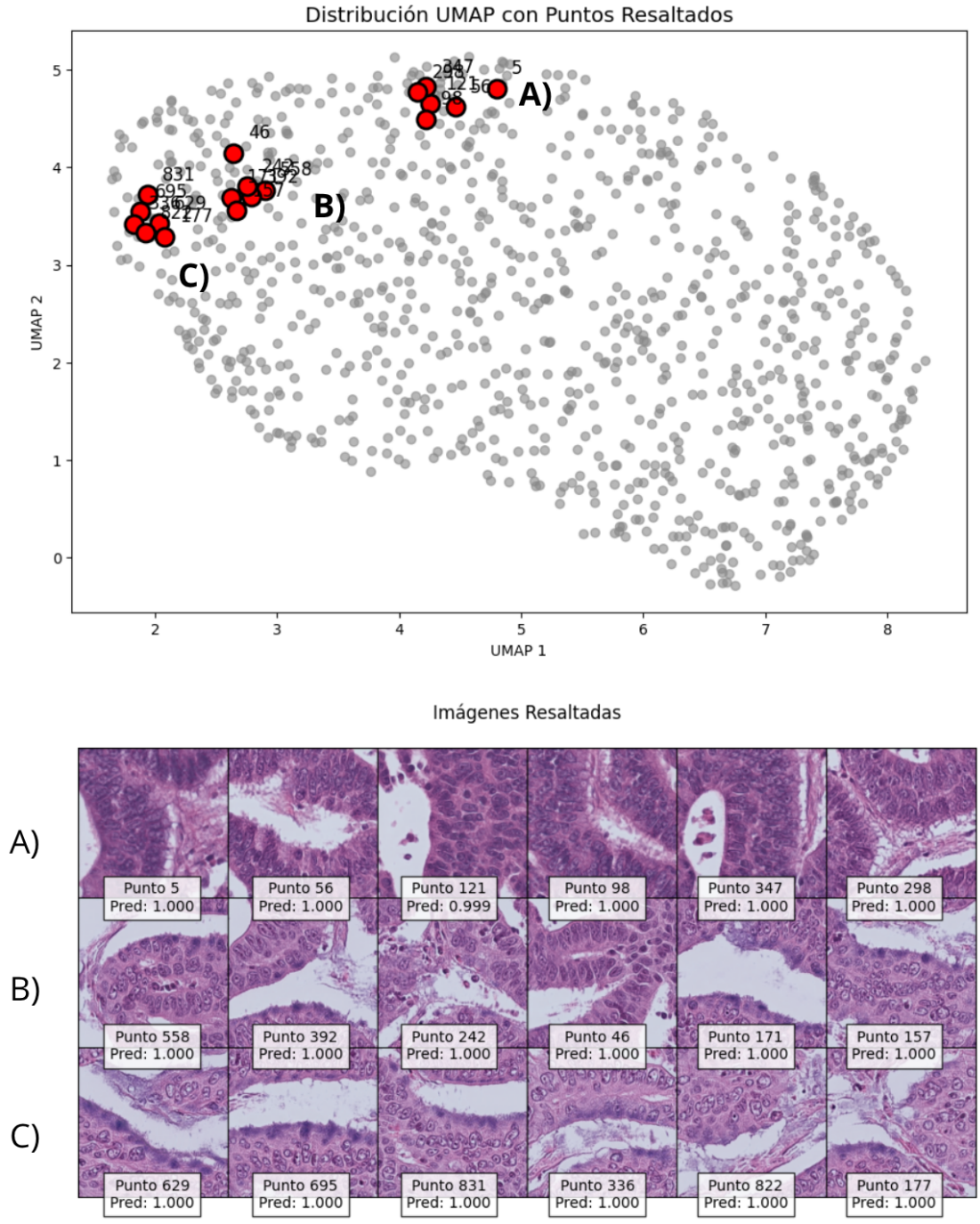


Figura 4.14: **Visualización del espacio latente de W mediante UMAP para la clase cancerosa, pero con el detalle de cada imagen.** Misma visualización del espacio latente de W mediante UMAP para la clase cancerosa, pero con el detalle de cada imagen (predicción y número de cada punto proyectado.)

El análisis de estas agrupaciones revela patrones distintivos asociados con un tejido de tipo canceroso; las imágenes en estas regiones muestran una pérdida de una arquitectura glandular típica, atipia nuclear marcada y núcleos hipercromáticos. El movimiento a través del espacio latente, de un clúster a otro, refleja cambios en estas características histopatológicas. Al igual que para el caso benigno, se presentan diferencias graduales entre cada uno de los clústers tanto en el color (más claro hacia C) como en una mayor desorganización o interacción con el estroma, aunque siempre manteniendo los rasgos característicos de una imagen de tipo cancerosa. La existencia de estos clústers diferenciados sugiere que el modelo ha aprendido a identificar y separar diferentes fenotipos visuales dentro de la categoría general de "cáncer". Cada fila de imágenes (A, B, C) confirma que el espacio latente está codificando características específicas, como las ya mencionadas anteriormente.

Discusión

La exploración del espacio latente W mediante UMAP, precedida por la confirmación de su superioridad sobre el espacio Z para la representación de características histopatológicas, ha proporcionado información valiosa sobre cómo la red organiza las representaciones internas. Los resultados demuestran que **el espacio latente W no es una mera distribución aleatoria, sino que posee una estructura semántica significativa**, con regiones de alta densidad que se correlacionan con fenotipos visuales coherentes.

La clara separación entre clases observada en la proyección UMAP del espacio W (Figura 4.9), en contraste con la del espacio Z, subraya la eficacia de la red de transformación para aprender un espacio de representación más desentrelazado. Dentro de este espacio W, la identificación de clústeres específicos para las clases benigna y cancerosa (Figuras 4.11 y 4.13) y la coherencia morfológica de las imágenes dentro de dichos clústeres indican que W codifica características histopatológicas específicas para cada clase. Las subagrupaciones (denotadas por letras) refuerzan esta idea, mostrando que incluso variaciones más sutiles dentro de un fenotipo general son capturadas y organizadas en vecindarios específicos del espacio latente.

La visualización de la densidad del espacio W (Figura 4.10) complementa este análisis al mos-

trar cómo se distribuyen globalmente las representaciones aprendidas. Las zonas de mayor densidad pueden interpretarse como las manifestaciones más típicas o frecuentes de las características que el modelo ha aprendido, tanto para la clase benigna como para la cancerosa. La estructura global del espacio, con regiones diferenciadas pero conectadas, sugiere que el modelo capta tanto la distinción entre clases como la variabilidad entre clases.

Estos hallazgos son de suma importancia desde la perspectiva de la explicabilidad. La capacidad de transformar regiones del espacio latente W a características morfológicas concretas abre una ventana hacia la comprensión de los criterios que el modelo de clasificación utiliza. El hecho de que W sea un espacio más desenredado facilita la interpretación de cómo las variaciones en este espacio se traducen en cambios visuales en las imágenes generadas o cómo podrían influir en las decisiones del clasificador. Se podría investigar si ciertas trayectorias o regiones en el espacio W son más influyentes para una clasificación particular, o incluso utilizar la estructura de W para guiar la generación de imágenes con características específicas, por ejemplo, para tareas de aumento de datos.

4.4. Resultados del Método Contrafactual

En esta sección se presentan los resultados obtenidos al aplicar el método de generación de contrafactuales descrito en la sección de metodología. El objetivo principal es demostrar la capacidad del enfoque para generar explicaciones visuales que ilustren cómo cambios específicos en las características de una imagen histopatológica pueden alterar la predicción de un clasificador preentrenado (DenseNet121), manteniendo al mismo tiempo la coherencia estructural global de la imagen original. Esto se logra mediante la exploración del espacio latente W del generador condicional, utilizando un conjunto fijo de tensores de ruido, como fue especificado con mayor detalle en la Sección 3.

Generación de un Ejemplo Contrafactual

Para ilustrar la capacidad del método, se seleccionó una imagen de entrada y se generó su correspondiente contrafactual. La Figura 4.15 muestra un ejemplo de este proceso.

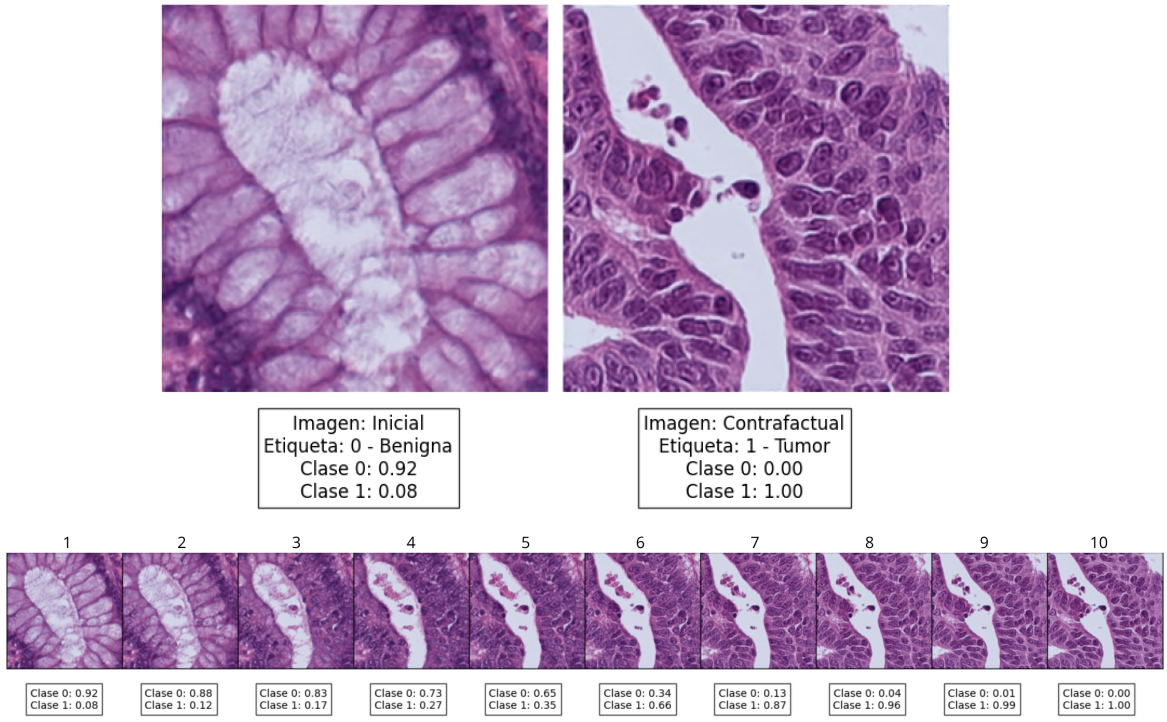


Figura 4.15: **Ejemplo de imagen inicial y su contrafactual generado por el método junto a la secuencia de interpolación entre ambas**. La transformación contrafactual se produce desde la clase benigna (0) a la clase cancerosa (1). En la secuencia de interpolación se observan los cambios graduales que conducen progresivamente a un cambio en la clasificación.

- **Imagen Inicial (Izquierda):** Corresponde a una muestra histopatológica que el clasificador DenseNet121 predice como benigna (Etiqueta: 0 - Benigna; Clase 0: 0,92 y Clase 1: 0,08). Visualmente, esta imagen presenta características consistentes con tejido benigno, como estructuras glandulares bien definidas y una celularidad relativamente uniforme con núcleos regulares.

- **Imagen Contrafactual (Derecha):** Esta imagen fue generada utilizando el vector latente w derivado de la imagen inicial, pero condicionando al generador con la etiqueta opuesta (Etiqueta: 1 - Tumor/Cancerosa). El clasificador predice esta imagen generada como cancerosa (Clase 0: 0,00 y Clase 1: 1,00).

Al comparar imágenes inicial y contrafactual (en la Figura 4.15), se observa que la estructura global y la disposición general del tejido se mantienen notablemente similares. Por ejemplo, la forma y orientación de las principales estructuras glandulares parecen conservadas. Esta conservación de la identidad de alto nivel es un resultado directo de la utilización de un conjunto fijo de tensores de ruido n durante la síntesis de ambas imágenes.

Sin embargo, se aprecian cambios sutiles pero significativos en las características morfológicas locales que son consistentes con la transformación de un fenotipo benigno a uno maligno. En la imagen contrafactual (derecha), se puede observar un incremento en la celularidad, mayor pleomorfismo nuclear (variación en tamaño y forma de los núcleos), hipercromasia (núcleos más oscuros), y una pérdida de la organización glandular regular. Son precisamente estas alteraciones en atributos de bajo y mediano nivel (textura, morfología celular) las que el clasificador interpreta como indicativas de malignidad, llevando al cambio en la predicción. Este ejemplo demuestra la capacidad del método para generar un contrafactual plausible que aísla las transformaciones visuales mínimas necesarias para invertir la decisión del clasificador.

Análisis de la Trayectoria Contrafactual mediante Interpolación Latente

Para obtener una comprensión más profunda de cómo los cambios visuales impactan progresivamente la decisión del clasificador, se aplicó la técnica de interpolación lineal en el espacio latente W entre el vector w_{orig} (correspondiente a la imagen inicial benigna) y el vector w_{cf} (utilizado para generar la imagen contrafactual cancerosa). La Figura 4.15 muestra una secuencia de imágenes generadas a lo largo de esta trayectoria contrafactual, junto con las predicciones del clasificador para cada paso de la interpolación. La secuencia, titulada “secuencia entre clases”, consta de 10 imágenes.

La trayectoria visualiza la transición gradual desde la apariencia benigna de la imagen original hasta la morfología cancerosa de la imagen contrafactual:

- **Extremo Inicial (Imagen 1, izquierda):** $p(\text{Clase } 0) = 0,92$ y $p(\text{Clase } 1) = 0,08$. Corresponde a la imagen original benigna.
- **Pasos Intermedios (Imágenes 2–5):** A medida que α aumenta (moviéndonos hacia la derecha), se observan cambios morfológicos incipientes. Los núcleos celulares comienzan a mostrar una ligera irregularidad y un aumento sutil en tamaño y coloración. Las predicciones del clasificador reflejan esta transición, con una disminución gradual en la probabilidad de la Clase 0 (benigna) y un aumento correlativo en la probabilidad de la Clase 1 (cancerosa). Por ejemplo: Imagen 5: $p(\text{Clase } 0) = 0,92$ y $p(\text{Clase } 1) = 0,08$.
- **Punto de Inflexión (Imagen 6):** $p(\text{Clase } 0) = 0,34$ y $p(\text{Clase } 1) = 0,66$. Aquí ocurre un cambio significativo: la predicción del clasificador se invierte, favoreciendo ahora la clase cancerosa. Visualmente, la imagen en este punto muestra un aumento más pronunciado de la atipia celular y una mayor desorganización de las estructuras glandulares.
- **Consolidación de la Clase Contrafactual (Imágenes 7–9):** Las características asociadas a la malignidad se vuelven progresivamente más evidentes.
- **Extremo Contrafactual (Imagen 10, derecha):** $p(\text{Clase } 0) = 0,00$ y $p(\text{Clase } 1) = 1,00$. Corresponde a la imagen contrafactual cancerosa.

De igual forma (en la Figura 4.16), podemos ver un ejemplo de transformación contrafactual pero desde una imagen etiquetada como cancerosa por el clasificador, hacia una imagen benigna:

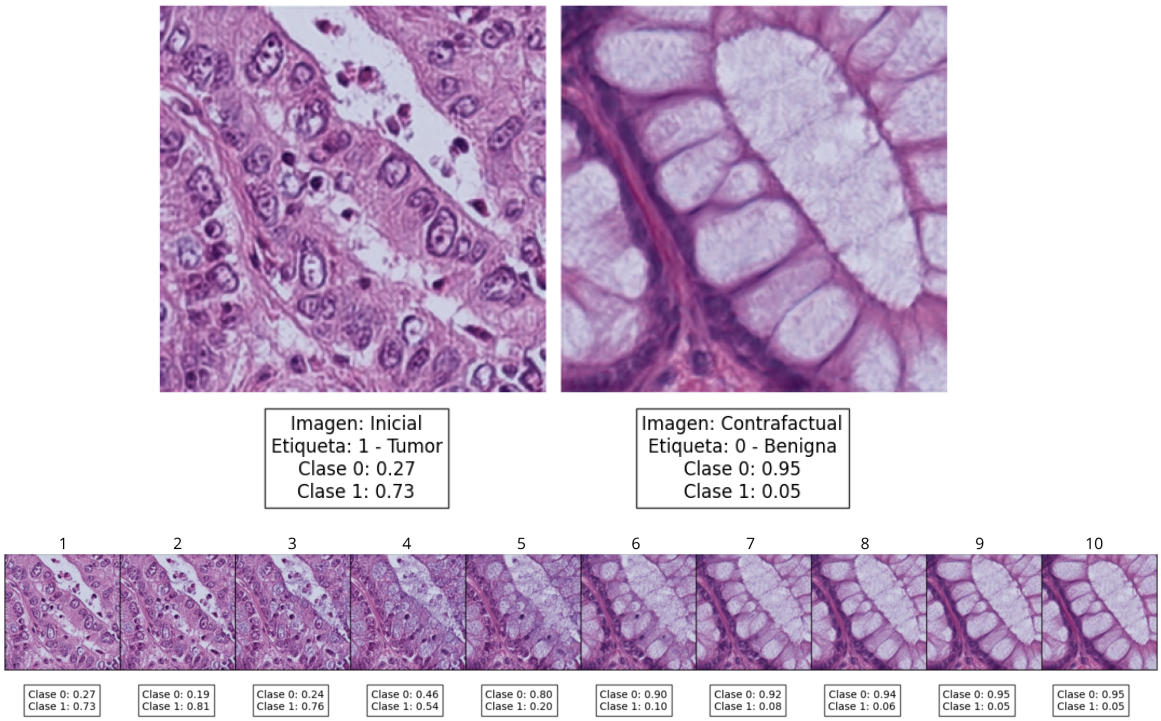


Figura 4.16: **Ejemplo de imagen inicial y su contrafactual generado por el método junto a la secuencia de interpolación entre ambas**. La transformación contrafactual se produce desde la clase cancerosa (1) a la clase benigna (0). En la secuencia de interpolación se observan los cambios graduales que conducen progresivamente a un cambio en la clasificación.

4.4.1. Diversidad en la Generación de Contrafactuales

Un aspecto fundamental de un método de explicación contrafactual robusto es su capacidad para generar no solo un único ejemplo, sino un conjunto diverso de contrafactuales. Esta diversidad es crucial porque pueden existir múltiples alteraciones mínimas en una imagen que conduzcan a un cambio en la predicción del clasificador, cada una resaltando diferentes sensibilidades del modelo.

La Figura 4.17 ilustra la riqueza del espacio latente W explorado por el generador. A la izquier-

da, se presenta una distribución UMAP de vectores latentes w, donde los puntos resaltados en azul y rojo indican muestras con diferentes predicciones del clasificador. A la derecha, se muestran las imágenes histopatológicas correspondientes a una selección de estos puntos, junto con la predicción del clasificador para cada una.

Esta figura demuestra que el generador es capaz de producir una amplia gama de apariencias histopatológicas con variaciones sutiles, incluso para puntos cercanos en el espacio UMAP. Por ejemplo:

- Puntos azules: Punto 24, Pred: 0,289; Punto 26, Pred: 0,283; Punto 39, Pred: 0,241. Estos tienden a mostrar características predominantemente benignas y reciben puntuaciones bajas del clasificador.
- Puntos rojos: Punto 31, Pred: 0,667; Punto 32, Pred: 0,710; Punto 47, Pred: 0,765. Estos exhiben rasgos más consistentes con la clase cancerosa y obtienen puntuaciones más altas.

La relevancia de esta figura para la diversidad contrafactual es la siguiente: si consideramos una imagen inicial representada por un punto azul (benigna), existen múltiples puntos rojos (malignos) en el espacio latente que podrían servir como sus contrafactuales. Cada uno de estos puntos rojos representa una instancia de imagen que:

- Es clasificada de forma diferente a la original (cancerosa en lugar de benigna).
- Mantiene una coherencia estructural con la base de la imagen original, debido al uso del mismo conjunto de tensores de ruido n si se partiera de un w base y se exploraran variaciones o se aplicara la etiqueta opuesta.
- Es visualmente distinta de otros posibles contrafactuales malignos. Por ejemplo, el “Punto 31” (Pred: 0,667) y el “Punto 47” (Pred: 0,765) son ambos clasificados como malignos pero presentan diferencias morfológicas notables.

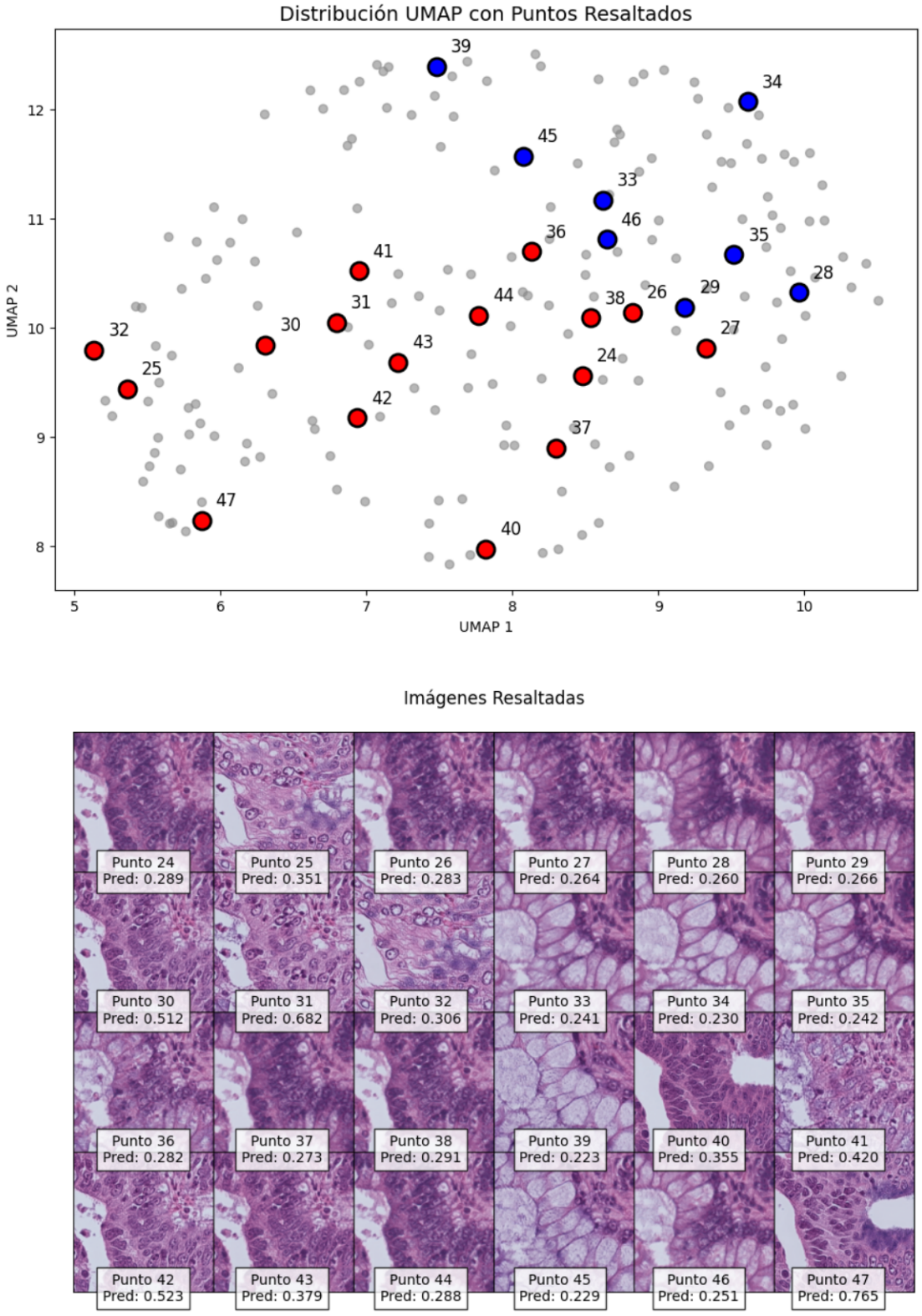


Figura 4.17: Diversidad de ejemplos contrafactuales generados a partir de distintos puntos en el espacio latente.

El método propuesto, al permitir la exploración del espacio latente $\mathcal{W}$ (a través de la variación de las semillas de entrada z_i que generan los w_i), facilita la identificación de esta diversidad. De hecho, para la creación misma de los contrafactuales se toma una imagen de entrada dada (o su vector w_{input}), y se proyectan en el espacio latente $\mathcal{W}$, una cantidad n vectores w_j condicionados con la etiqueta opuesta $\bar{\ell}_{\text{input}}$. De esta manera se producen imágenes I_j clasificadas como $\bar{\ell}_{\text{input}}$ y se selecciona la que tenga el puntaje de clasificación más cercano a la clase objetivo, aunque esto podría ser claramente un hiperparámetro.

La exploración de múltiples contrafactuales diversos para una misma imagen de entrada proporciona una explicación más completa del comportamiento del clasificador. En lugar de señalar una única característica o cambio, la diversidad permite entender que el modelo puede ser sensible a diferentes combinaciones de atributos o a diferentes manifestaciones de una misma patología. Esto es particularmente valioso en el dominio histopatológico, donde la heterogeneidad de las enfermedades es común.

Discusión e Implicaciones para la Explicabilidad

Los resultados obtenidos mediante la generación de contrafactuales, el análisis de trayectorias de interpolación y la consideración de su diversidad, ofrecen herramientas potentes para la interpretabilidad de los modelos de clasificación en histopatología.

La capacidad de generar imágenes contrafactuales que difieren mínimamente de la original, pero que provocan un cambio en la predicción del clasificador, permite identificar visualmente las características a las que el modelo es sensible. El uso de ruido fijo es fundamental, asegurando que las diferencias se deban a cambios semánticos. La interpolación latente descompone la transición entre clases, permitiendo visualizar la frontera de decisión e identificar características conductoras.

La consideración de la diversidad contrafactual, evidenciada por la variedad de muestras en el espacio latente explorado (ver Figura 4.17), enriquece aún más la explicación. Revela que no hay una única “razón” para un cambio de clasificación, sino potencialmente múltiples caminos

o combinaciones de características que el modelo considera relevantes. Esta perspectiva multifacética es esencial para una comprensión profunda y para la validación de modelos en entornos médicos e histopatológico en particular, donde las decisiones deben ser robustas y bien fundamentadas.

4.5. Resultados de la evaluación con expertos

Perfil de los expertos evaluadores

Para asegurar una evaluación sólida y multidisciplinaria del método propuesto, participaron un total de **13 profesionales** con formación en disciplinas relacionadas con la histopatología (ver Tabla 4.6). Las profesiones incluyen, entre otros, tecnología médica, bioquímica, técnico especialista en anatomía patológica, técnico de laboratorio y médico. Esta diversidad de perfiles, que abarca desde la preparación de muestras en laboratorio hasta el diagnóstico clínico microscópico, aporta múltiples perspectivas al análisis.

Asimismo, el grupo posee un nivel de experiencia promedio de 11,23 años en análisis histopatológico y diagnóstico de tejidos, lo cual refuerza la fiabilidad de sus juicios en las cuatro etapas del experimento.

La heterogeneidad profesional y trayectoria acumulada de los participantes expertos garantizan que las evaluaciones reflejen tanto la fidelidad técnica como la validez clínica de las imágenes presentadas.

Resultados Etapa 1: Evaluación de Imágenes Sintéticas

En la Etapa 1 se presentó de forma aleatoria un total de 20 imágenes (mismo conjunto para todos), 10 reales y 10 sintéticas, a cada uno de los 13 expertos, pidiéndoles clasificar cada imagen como *real*, *sintética* o *indistinguible*. El objetivo de esta fase fue cuantificar el grado de realismo percibido en las imágenes generadas por el modelo y su capacidad de engañar al observador

Profesión	**Cantidad**
Tecnólogo médico	3
Bioquímica	2
Técnico especialista anatomía patológica	2
Técnico anatomía sanitario	1
Técnico de laboratorio	1
Bióloga	1
Microbióloga	1
Médico	1
Biomédico	1
Participantes totales	13
Años de experiencia promedio	11,23

Tabla 4.6: **Estadísticas de los usuarios expertos encuestados.** Se muestra la profesión de cada uno de los expertos encuestados, junto con los años de experiencia promedio.

experto.

	Respuestas		
Categoría original	**Real**	**Sintética**	**Indistinguible**
Reales	80 % (104)	10 % (13)	10 % (13)
Sintéticas	60 % (78)	30 % (39)	10 % (13)

Tabla 4.7: **Respuestas de la Etapa 1 del experimento (Sección 3.7.1).** Corresponde al porcentaje de respuestas que todos los usuarios contestaron como Real, Sintéticas o Indistinguibles, según las 20 imágenes originalmente etiquetadas como Reales y Sintéticas (10 por cada una).

Interpretación de resultados

- **Imágenes reales:** el 80 % de las veces fueron correctamente identificadas como reales, mientras que solo 10 % fueron etiquetadas erróneamente como sintéticas y otro 10 % como indistinguibles. Esto confirma que los evaluadores reconocen mayoritariamente las texturas y patrones propios de las imágenes obtenidas del dataset original.
- **Imágenes sintéticas:** un 60 % de las muestras fueron clasificadas como *reales* y un 10 % fueron clasificadas como *indistinguibles* lo que sumadas corresponden a un *fool rate* inclusivo del 70 %. El 30 % restante se identificó correctamente como sintético.
- **Accuracy:** el valor de la precisión global queda como:

$$\text{Accuracy} = \frac{104 + 39}{260} \approx 0{,}55 \;=\; 55\ \%.$$

Discusión Estos resultados indican que las imágenes generadas por el modelo poseen un nivel de realismo notable, capaz de engañar a los expertos en 7 de cada 10 casos. La alta tasa de detección correcta de las imágenes reales (80 %) avala la validez del test y descarta una posible tendencia de los evaluadores a responder al azar. Del mismo modo, una precisión global muy cercana a 0,50 indica que los expertos no distinguen sistemáticamente las imágenes reales de las sintéticas, lo cual evidencia un alto nivel de realismo en las muestras generadas.

Esta etapa demuestra que el método de generación alcanza un realismo suficiente para confundir al ojo clínico en la mayoría de los casos, cumpliendo el objetivo de evaluar la *indistinguibilidad* de las muestras sintéticas respecto a las reales. Estos hallazgos motivan un análisis más detallado de las características específicas que facilitan o dificultan el engaño, así como la comparación con otras técnicas de síntesis en futuras líneas de trabajo.

Resultados Etapa 2: Generación y evaluación de Imágenes Contrafactuales

En esta etapa se calculó, para cada par imagen original–contrafactual, el índice compuesto S (ver Sección 3.7.1), de modo que valores altos de S indican mayor presencia agregada de las cinco características patológicas. Recordemos que nos interesa conocer si en los contrafactua-

les generados existe o no existe la presencia de las 5 características predefinidas de antemano, dependiendo si el contrafactual es canceroso o benigno, respectivamente (las características se detallan en la Sección 3.7.1). La Tabla 4.8 resume los resultados obtenidos en los 10 pares de imágenes evaluadas por los expertos (Las imágenes se muestran en el Anexo 5).

Imagen	**Clase**	**S** (%)
01 (Figura C.11)	Benigno (0)	68,40
02 (Figura C.12)	Benigno (0)	67,36
03 (Figura C.13)	Canceroso (1)	64,93
04 (Figura C.14)	Benigno (0)	77,43
05 (Figura C.15)	Benigno (0)	51,04
06 (Figura C.16)	Benigno (0)	29,51
07 (Figura C.17)	Canceroso (1)	69,44
08 (Figura C.18)	Canceroso (1)	81,75
09 (Figura C.19)	Canceroso (1)	73,02
10 (Figura C.20)	Canceroso (1)	86,11

Tabla 4.8: Valores de S por imagen y clase de contrafactual.

A continuación se evaluó la capacidad discriminativa de S para distinguir contrafactuales benignos (0) de cancerosos (1), usando un umbral de decisión $S \geq 50$ %.

Matriz de confusión y métricas de clasificación

$$\begin{pmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{pmatrix} = \begin{pmatrix} 1 & 4 \\ 0 & 5 \end{pmatrix}$$

$$\text{Sensibilidad (TPR)} = \frac{5}{5} = 1{,}00 \quad \text{Especificidad (TNR)} = \frac{1}{1+4} = 0{,}20$$

Interpretación

- El **índice promedio** para contrafactuales benignos resultó $\bar{S}_{\text{benigno}} \approx 58{,}8\,\%$ frente a $\bar{S}_{\text{canceroso}} \approx 75{,}0\,\%$.
- Con el umbral fijado en 50 %, sólo 1 de las 5 imágenes benignas (TN) quedó correctamente clasificada (Imagen 06), mientras que las 5 cancerosas fueron detectadas sin error (TPR = 100 %).
- La **especificidad baja** (20 %) refleja un elevado número de falsos positivos: 4 contrafactuales benignos superaron el umbral y se etiquetaron como cancerosos.

Discusión La perfecta sensibilidad sugiere que el índice S capta correctamente las características definitorias de los contrafactuales cancerosos, cumpliendo el objetivo de detectar la presencia de rasgos patológicos. Sin embargo, la baja especificidad revela que los contrafactuales benignos también exhiben, en promedio, valores elevados de S (aunque menores que los cancerosos). Esto puede deberse a la aparición residual de rasgos patológicos en la generación o a un umbral demasiado conservador para distinguir ambas clases.

Resultados Etapa 3: Evaluación de Consistencia Visual en Imágenes Reconstruidas

En la Etapa 3 se presentaron a los expertos los 10 pares de imagen original–reconstruida para valorar (1) la conservación del patrón biológico y (2) el realismo visual de las reconstrucciones, según la escala de Likert 1–5.

Interpretación de resultados

- En el 56 % de las valoraciones, las reconstrucciones mantuvieron completamente los patrones biológicos de la imagen original, lo que indica una elevada fidelidad del encoder para preservar las características morfológicas relevantes.
- Un 38 % de las respuestas apuntaron a una pérdida parcial de dichos patrones; esto es

consistente con la compresión de la información en el espacio latente, donde detalles finos pueden atenuarse sin desaparecer por completo.

- Sólo un 6 % de las valoraciones consideró que el patrón se perdió completamente, un porcentaje reducido que sugiere que los fallos totales de reconstrucción son poco frecuentes.
- La puntuación media de realismo visual (4,12 de 5) revela que, pese a ciertas alteraciones menores, las imágenes reconstruidas conservan un alto grado de verosimilitud clínica y estética, superando ampliamente la mitad del rango máximo.

Respuesta	**% (n)**	**Interpretación**
Se mantiene	56 % (73)	Conservación completa del patrón
Se pierde parcialmente	38 % (49)	Pérdida parcial de estructuras
Se pierde completamente	6 % (8)	Ruptura total del patrón
Realismo visual promedio		4.12 (escala 1–5)

Tabla 4.9: **Respuestas de la Etapa 3 del experimento (Sección 3.7.1).** Corresponde al porcentaje y cantidad de respuestas en 3 categorías, según los 10 pares de imágenes original-reconstruida mostradas al usuario. Además, se muestra el realismo visual promedio.

Discusión Estos hallazgos confirman que el mecanismo de codificación–decodificación utilizado logra un equilibrio óptimo entre compresión y preservación de la información histopatológica:

La mayoría de las reconstrucciones (56 %) son prácticamente indistinguibles de las originales en cuanto a patrón biológico, lo que hace útil la aplicabilidad del modelo en tareas donde se requiera mantener integridad morfológica.

La pérdida parcial (38 %) se sitúa dentro de lo esperado para autoencoders de alta dimensionalidad, y podría mitigarse ajustando la capacidad del espacio latente o empleando términos de regularización específicos que penalicen la distorsión de texturas críticas.

El bajo porcentaje de fallos completos (6 %) sugiere robustez general, pero invita a investigar casos puntuales para detectar si ciertas estructuras (por ejemplo, zonas de alta variabilidad celular) son más susceptibles a errores de reconstrucción.

El alto realismo visual promedio (4,12) subraya que, incluso cuando existe pérdida parcial de patrones, la apariencia global se mantiene coherente con las expectativas de un observador experto.

Resultados Etapa 4: Utilidad del Mecanismo

En la Etapa 4 se evaluó la percepción de los expertos sobre la utilidad de visualizar contrafactuales para el diagnóstico médico (Pregunta 1) y, específicamente, en casos difíciles o equívocos (Pregunta 2). Cada experto respondió ambas preguntas utilizando una escala de Likert de 1 (*Nada útil*) a 5 (*Extremadamente útil*).

Respuesta (Likert)	**Pregunta 1**	**Pregunta 2**
5 (Extremadamente útil)	23 % (3)	38,5 % (5)
4 (Muy útil)	69,3 % (9)	53,8 % (7)
3 (Moderadamente útil)	0 %	0
2 (Poco útil)	0 %	0
1 (Nada útil)	7,7 % (1)	7,7 % (1)

Tabla 4.10: **Respuestas de la Etapa 4 del experimento (Sección 3.7.1).** Corresponde al porcentaje y cantidad de respuestas en una escala de Likert (1-5), según las 2 preguntas finales realizadas al usuario.

Interpretación de resultados

- En la **Pregunta 1**, un **92,3 %** de los expertos (12/13) calificaron la visualización de contrafactuales como *muy* o *extremadamente útil* para reforzar su decisión diagnóstica. Solo un evaluador (7,7 %) la consideró *nada útil*.
- En la **Pregunta 2**, el **92,3 %** de los expertos (12/13) también calificaron como *muy* o *extremadamente útil* la herramienta en contextos de casos difíciles o intermedios.
- No se registraron respuestas en las categorías intermedias (3 o 2), lo que evidencia una valoración polarizada y mayoritariamente positiva.

Discusión Los resultados muestran un respaldo consistente de los expertos hacia la integración de contrafactuales en el flujo diagnóstico:

La *casi unánime* calificación alta (4–5) en ambas preguntas indica que el mecanismo cumple su objetivo principal: ofrecer un soporte visual que fortalece la confianza del especialista en sus juicios microscópicos.

El incremento relativo de evaluaciones *extremadamente útiles* en la Pregunta 2 (38,5 % vs. 23 %) sugiere que los contrafactuales aportan un valor adicional en escenarios complejos, donde la distinción entre tejido benigno y maligno es más sutil o complejo de discriminar sin herramientas de inteligencia artificial.

La ausencia de valoraciones neutrales o bajas (3 o 2) refuerza la percepción de que la herramienta no pasa desapercibida; impacta claramente en la interpretación clínica. Solo un caso de valoración negativa en cada pregunta podría corresponder a diferencias personales en estilo de trabajo o a una curva de aprendizaje de la interfaz.

Finalmente, la Etapa 4 confirma que los expertos no solo reconocen la plausibilidad de los contrafactuales, sino que los perciben como un recurso de gran utilidad para mejorar la toma de decisiones diagnósticas, especialmente en aquellos casos de alta complejidad. Como siguiente paso, se recomienda explorar la integración de estos resultados en un prototipo de software clínico y evaluar su impacto en métricas de rendimiento diagnóstico como el tiempo total del mismo,

el incremento en la tasa de diagnósticos correctos o la capacidad para resolver casos difíciles o equívocos.

Comparación con el Método de Referencia: Chexplaining in Style

Para evaluar la efectividad del método contrafactual propuesto, se realizó una comparación directa con la implementación de "Chexplaining in Style". A fin de garantizar una comparación justa y centrada en las estrategias de generación de contrafactuales, ambos métodos partieron de los mismos componentes pre-entrenados: el generador StyleGAN2-ADA, el codificador y el clasificador, todos entrenados sobre el dataset NCT-CRC-HE-100K. Esta aproximación permite aislar y evaluar las diferencias en los algoritmos de manipulación latente, en lugar de las arquitecturas de base. La comparación se centró en tres ejes principales: eficiencia computacional, calidad de las explicaciones contrafactuales y aplicabilidad al dominio de la histopatología.

Existe una diferencia fundamental en la forma en que cada método busca los contrafactuales, lo que impacta directamente en la eficiencia computacional. El método propuesto se basa en una exploración del espacio latente mediante muestreo. Para una imagen de entrada, se generan N vectores latentes w condicionados con la etiqueta opuesta y, a partir de ellos, N imágenes candidatas en un solo proceso de generación. Este proceso es inherentemente paralelizable y no requiere optimización iterativa, lo que lo hace computacionalmente eficiente. En contraste, "Chexplaining in Style" se basa en un proceso de optimización iterativo. Primero, requiere un muestreo extenso del espacio latente para realizar un Análisis de Componentes Principales (PCA) e identificar las direcciones de máxima varianza. Luego, para cada imagen, formula la búsqueda del contrafactual como un problema de optimización que busca minimizar los coeficientes de manipulación α_i a lo largo de estas direcciones, lo cual exige múltiples ciclos de refinamiento, dependiendo del número de coeficientes (por defecto son 20). En resumen, el método propuesto es significativamente más rápido, ya que reemplaza un costoso bucle de optimización iterativa por un proceso de muestreo y generación directa. Como se muestra en la Tabla 4.11, para la generación de 50 contrafactuales, el método propuesto tarda 3,1 segundos, mientras que el método "Chexplaining in Style" tarda 224,7 segundos.

Método	**n° de muestras**	**n° de α_i**	**Tiempo**
Propuesto	50	-	3,1 s
Chexplaining in Style	50	20	224,7 s

Tabla 4.11: **Tabla de Comparación de tiempo entre métodos.** Se compararon ambos métodos para la generación de 50 imágenes histopatológicas y sus respectivos contrafactuales.

En cuanto a la calidad de las explicaciones, se evaluó su interpretabilidad, diversidad y la relevancia de los cambios visuales. El método propuesto permite generar un conjunto diverso de contrafactuales para una única imagen de entrada al explorar distintas regiones del espacio latente (como se pudo ver en la Figura 4.17). Esta diversidad es crucial, ya que refleja que múltiples combinaciones de atributos pueden llevar a un cambio de clasificación, ofreciendo una explicación más completa y robusta. Adicionalmente, el uso de un vector de ruido fijo (n) para todas las generaciones asegura que la estructura global del tejido se mantenga, atribuyendo los cambios en la clasificación únicamente a variaciones semánticas como la morfología celular o la textura. La técnica de interpolación latente también ofrece una trayectoria contrafactual que visualiza de forma suave y continua la transición entre clases, permitiendo identificar de manera gradual los cambios visuales que impactan la decisión del clasificador. Por su parte, "Chexplaining in Style" está diseñado para encontrar la mínima modificación necesaria para cambiar la predicción a lo largo de las direcciones semánticamente más relevantes, que son las componentes principales. Mientras que este enfoque produce una explicación única y optimizada, el método propuesto genera un conjunto de explicaciones más diverso y dinámico, con la ventaja distintiva de visualizar una transición continua mediante interpolación para mejorar la interpretabilidad.

Además, como se muestra en la Figura 4.18, el método "Chexplaining in Style" no logró generar contrafactuales que inducieran un cambio de clase. Los cambios generados, además, no presentan una coherencia semántica clara; de hecho, los contrafactuales resultantes parecen imágenes completamente nuevas, sin relación estructural evidente con las originales.

Esta deficiencia se debe a que dicho método no impone restricciones al vector latente $\mathbf{w}$ durante la manipulación, lo que puede conducir a modificaciones arbitrarias de la estructura. En contraste, el método propuesto en esta tesis incorpora restricciones informadas por un análisis previo, lo que permite preservar las características morfológicas relevantes de las imágenes histopatológicas durante la generación contrafactual.

Cabe destacar que "Chexplaining in Style" no es un enfoque incorrecto. Sin embargo, su efectividad está fuertemente ligada al dominio de aplicación. En el caso de las radiografías de tórax, las imágenes presentan una estructura anatómica fija (como por ejemplo, el esqueleto) y una escala de grises homogénea, lo que reduce la necesidad de preservar manualmente la coherencia estructural. En cambio, en el dominio histopatológico, donde existe una alta variabilidad morfológica y cromática, la preservación explícita de la estructura es fundamental para generar explicaciones útiles y clínicamente válidas.

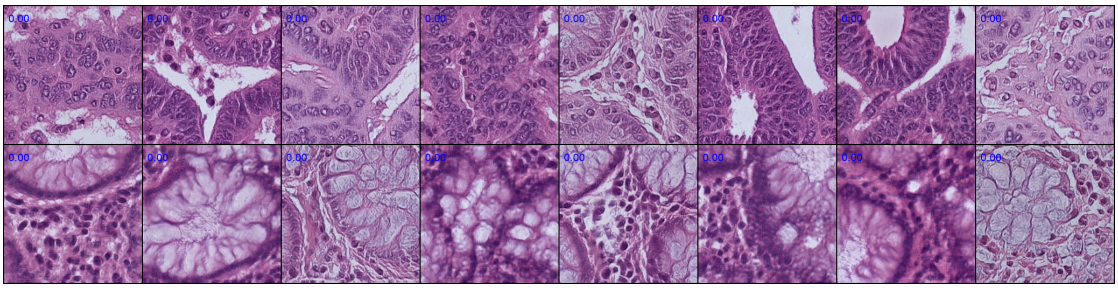


Figura 4.18: **Ejemplo de generación de contrafactuales del método Chexplaining in Style.** La configuración del método para esta generación es para 50 muestras pero solo se muestran las mejores 8. Los puntajes de clasificación se muestran en la parte superior izquierda de cada imagen. La primera fila indica las imágenes generadas al azar por el método generativo y la segunda fila indica los contrafactuales hallados para cada una de ellas.

Finalmente, la aplicabilidad de cada método se consideró frente a los desafíos del análisis de imágenes histopatológicas. El método propuesto utiliza StyleGAN2-ADA, una elección deliberada por su capacidad para manejar conjuntos de datos escasos y con alta variabilidad, una

característica común en histopatología. La generación de contrafactuales diversos se alinea bien con la naturaleza heterogénea de las patologías en este dominio, y la estrategia de ruido fijo preserva la arquitectura morfológica global mientras permite cambios en características locales y texturales, que son cruciales para el diagnóstico. “Chexplaining in Style”, en cambio, fue originalmente diseñado para radiografías de tórax, donde los cambios relevantes son a gran escala. Su dependencia de PCA para identificar la varianza global podría ser menos sensible a las transformaciones sutiles y texturales que definen la patología en imágenes de tejidos histopatológicos. Aunque incorpora restricciones de plausibilidad médica, estas fueron concebidas para estructuras anatómicas, y su adaptación al tejido es un desafío. Por todo ello, el método propuesto demuestra una mayor aplicabilidad nativa al dominio histopatológico, con un diseño mejor alineado con la naturaleza heterogénea de los datos.

Capítulo 5

Conclusiones

Este trabajo de tesis se centró en el desarrollo y la evaluación de un sistema generativo avanzado para la generación y manipulación de imágenes histopatológicas, con un énfasis particular en la creación de explicaciones contrafactuales visuales. Los resultados obtenidos a lo largo de las diferentes etapas experimentales permiten extraer conclusiones significativas y plantear prometedoras líneas de investigación futura.

Conclusiones Principales

Se demostró que una estrategia de entrenamiento en dos etapas —iniciando con un preentrenamiento enfocado exclusivamente en la reconstrucción de imágenes y luego introduciendo la tarea de clasificación— es crucial para optimizar el rendimiento del sistema. Este enfoque resultó en una mejora sustancial en las métricas de reconstrucción (LPIPS, PSNR, SSIM) y en la calidad perceptual de las imágenes generadas (menor FID) en comparación con un entrenamiento conjunto desde el inicio. Esto subraya la importancia de permitir que el codificador aprenda una inversión robusta al espacio latente W antes de ser influenciado por los gradientes de la tarea de clasificación.

El análisis sobre el balance de la función de pérdida mostró una inevitable relación de compromiso entre la fidelidad a nivel de píxel, la calidad perceptual y la consistencia en el espacio

latente. Priorizar la fidelidad numérica tiende a generar imágenes más borrosas, mientras que una mayor ponderación de la consistencia latente mejora la estabilidad en el espacio W a costa de una ligera reducción visual. En este contexto, la configuración identificada como "Baseline Equilibrado" ofreció un buen compromiso entre estos factores, convirtiéndose en la alternativa más práctica para esta investigación.

La exploración del espacio latente confirmó la superioridad del espacio W sobre el espacio Z en términos de desentrelazamiento y separabilidad de clases (benigna vs. cancerosa), visualizada mediante UMAP. El espacio W no solo organiza las muestras de manera coherente con su clasificación, sino que también estructura la variabilidad entre clases en clústeres semánticamente significativos. Se observó que regiones específicas y densas en W corresponden a fenotipos visuales distintivos y coherentes dentro de cada clase, capturando variaciones sutiles en la morfología celular, textura y coloración. Esto es fundamental para la interpretabilidad, ya que vincula regiones del espacio latente con características histopatológicas concretas.

El método propuesto para la generación de contrafactuales demostró ser eficaz en la creación de imágenes que alteran la predicción de un clasificador preentrenado ($DenseNet121$) mediante modificaciones visuales mínimas y localizadas, manteniendo la estructura global de la imagen original. La clave de esta coherencia estructural radica en el uso de un conjunto fijo de tensores de ruido durante la generación de imágenes. Las transformaciones visuales observadas en los contrafactuales (e.g., de benigno a canceroso) se alinean con cambios histopatológicos conocidos, como el aumento de la actividad mitótica, pleomorfismo nuclear y celular, y arquitectura glandular alterada.

En la comparación con el método "Chexplaining in Style", el enfoque desarrollado en esta tesis demostró una clara superioridad tanto en eficiencia como en validez semántica. Mientras el método de referencia tardó 224,7 segundos en generar 50 contrafactuales, el sistema propuesto logró la misma tarea en apenas 3,1 segundos. Además, a diferencia del método comparado, el modelo fue capaz de inducir cambios de clase coherentes en imágenes histopatológicas.

La interpolación lineal en el espacio latente W entre una imagen original y su contrafactual permitió visualizar trayectorias de transformación graduales. Estas trayectorias no solo ilustran

los cambios morfológicos progresivos, sino que también revelan cómo evoluciona la puntuación del clasificador, permitiendo identificar "puntos de inflexión" donde la predicción cambia de clase. Esta técnica es una herramienta poderosa para entender qué conjunto de características visuales son suficientes para cruzar la frontera de decisión del clasificador.

Se destacó que para una imagen dada, pueden existir múltiples contrafactuales visualmente distintos que provocan el mismo cambio de clasificación. Esta diversidad, facilitada por la exploración del espacio latente W, proporciona una explicación más rica y completa, sugiriendo que el clasificador puede ser sensible a diferentes combinaciones de atributos o manifestaciones patológicas.

La evaluación con 13 expertos en el área arrojó resultados muy positivos:

- **Alto Realismo de Imágenes Sintéticas:** Las imágenes generadas por el modelo lograron un "fool rate" inclusivo del 70 % con los expertos, indicando un notable nivel de realismo.
- **Consistencia de Reconstrucciones:** El 56 % de las reconstrucciones mantuvieron completamente los patrones biológicos originales, y el realismo visual promedio fue de 4,12 con un máximo de 5.
- **Pertinencia de Contrafactuales:** Aunque el índice S para contrafactuales benignos mostró una especificidad baja (20 %), la sensibilidad para detectar características patológicas en contrafactuales cancerosos fue perfecta (100 %), sugiriendo que el método capta rasgos definitorios de malignidad.
- **Alta Utilidad Percibida:** Un 92,3 % de los expertos consideró la visualización de contrafactuales como "muy útil" o "extremadamente útil" para reforzar decisiones diagnósticas, especialmente en casos difíciles o equívocos.

Perspectivas Futuras

Los resultados y conclusiones de esta tesis abren varias líneas de investigación:

La baja especificidad del índice S en la Etapa 2 de la evaluación con expertos sugiere que los

contrafactuales benignos generados aún pueden exhibir características que los expertos asocian con patología, o que el umbral de S necesita un ajuste más fino. Futuros trabajos podrían enfocarse en refinar el proceso de generación o selección de contrafactuales benignos para asegurar una mayor "limpieza"de rasgos patológicos residuales, o explorar métricas de evaluación alternativas y adaptativas.

Si bien se demostró la existencia de diversidad, se podría investigar métodos más explícitos para generar conjuntos de contrafactuales que cubran diferentes “estrategias” de cambio de clase. Esto podría incluir algoritmos de búsqueda en el espacio latente que optimicen no solo el cambio de clase sino también la diferencia visual entre los contrafactuales generados, manteniendo la plausibilidad.

Como sugieren los resultados de la evaluación con expertos, existe un claro interés por parte de los profesionales en utilizar estas herramientas. Un paso lógico sería el desarrollo de un prototipo de software clínico que integre la generación de contrafactuales y la visualización de trayectorias. Dicha herramienta podría permitir a los patólogos explorar interactivamente el impacto de modificar ciertas características y observar cómo responde el clasificador, mejorando la confianza y la comprensión del modelo de IA.

El presente trabajo se centró en una clasificación binaria (benigno vs. canceroso). Una extensión natural sería aplicar y adaptar estos métodos a problemas de clasificación multiclase (e.g., diferentes subtipos de cáncer) o tareas para cuantificar el grado de “cancerosidad”. Esto presentaría nuevos desafíos en la definición y generación de contrafactuales entre múltiples estados.

Sería valioso realizar estudios para cuantificar el impacto real de estas herramientas de xAI en el rendimiento diagnóstico de los patólogos, midiendo métricas como la precisión, el tiempo de diagnóstico, la concordancia interobservador y la confianza en la decisión, especialmente en casos ambiguos.

Conclusión General

En resumen, esta tesis demostró la utilidad de usar modelos generativos avanzados en la creación de explicaciones contrafactuales para imágenes histopatológicas. Los resultados son alentadores tanto desde la perspectiva técnica de la calidad de generación y manipulación, como desde la perspectiva de la utilidad clínica percibida por los expertos. Las futuras líneas de investigación propuestas tienen el potencial de refinar aún más estas herramientas y facilitar su transición hacia la práctica clínica, contribuyendo a una inteligencia artificial más transparente, confiable y colaborativa en el campo de la histopatología.

Bibliografía

[Abdal et al., 2020] Abdal, R., Qin, Y., and Wonka, P. (2020). Image2stylegan++: How to edit the embedded images?

[Adebayo et al., 2020] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2020). Sanity checks for saliency maps.

[Agarwal and Nguyen, 2020] Agarwal, C. and Nguyen, A. (2020). Explaining image classifiers by removing input features using generative models.

[Aksac et al., 2019] Aksac, A., Demetrick, D., Ozyer, T., and Alhajj, R. (2019). Brecahad: a dataset for breast cancer histopathological annotation and diagnosis. *BMC Research Notes*, 12.

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan.

[Atad et al., 2022] Atad, M., Dmytrenko, V., Li, Y., Zhang, X., Keicher, M., Kirschke, J., Wiestler, B., Khakzar, A., and Navab, N. (2022). Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan.

[Atad et al., 2024] Atad, M., Schinz, D., Moeller, H., Graf, R., Wiestler, B., Rueckert, D., Navab, N., Kirschke, J. S., and Keicher, M. (2024). Counterfactual explanations for medical image classification and regression using diffusion autoencoder. *Machine Learning for Biomedical Imaging*, 2(iMIMIC 2023):2103–2125.

[Bach et al., 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise

relevance propagation. *PLOS ONE*, 10(7):1–46.

[Baumgartner et al., 2018] Baumgartner, C. F., Koch, L. M., Tezcan, K. C., Ang, J. X., and Konukoglu, E. (2018). Visual feature attribution using wasserstein gans.

[Bera et al., 2019] Bera, Kaustav, S., Kurt A., R., David L., V., Vamsidhar, M., and Anant (2019). Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*.

[Bigolin Lanfredi et al., 2019] Bigolin Lanfredi, R., Schroeder, J. D., Vachet, C., and Tasdizen, T. (2019). *Adversarial Regression Training for Visualizing the Progression of Chronic Obstructive Pulmonary Disease with Chest X-Rays*, page 685–693. Springer International Publishing.

[Bojarski et al., 2018] Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Ackel, L. J., Muller, U., Yeres, P., and Zieba, K. (2018). Visualbackprop: Efficient visualization of cnns for autonomous driving. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4701–4708.

[Bulten et al., 2019] Bulten, W., Bándi, P., Hoven, J., Loo, R. v. d., Lotz, J., Weiss, N., Laak, J. v. d., Ginneken, B. v., Hulsbergen-van de Kaa, C., and Litjens, G. (2019). Epithelium segmentation using deep learning in he-stained prostate specimens with immunohistochemistry as reference standard. *Scientific reports*, 9(1):864.

[Chang et al., 2019] Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. (2019). Explaining image classifiers by counterfactual generation.

[Chattopadhay et al., 2018] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847.

[Chen et al., 2019] Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. (2019). Isolating sources of disentanglement in variational autoencoders.

[Chen et al., 2016] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets.

[Choi et al., 2018] Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation.

[Choi et al., 2020] Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains.

[Chong et al., 2021] Chong, M. J., Chu, W.-S., Kumar, A., and Forsyth, D. (2021). Retrieve in style: Unsupervised facial feature transfer and retrieval.

[Chuquicusma et al., 2018] Chuquicusma, M. J. M., Hussein, S., Burt, J., and Bagci, U. (2018). How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 240–244.

[Collins et al., 2020] Collins, E., Bala, R., Price, B., and Süsstrunk, S. (2020). Editing in style: Uncovering the local semantics of gans.

[Dabkowski and Gal, 2017] Dabkowski, P. and Gal, Y. (2017). Real time image saliency for black box classifiers.

[Dastin, 2018] Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

[Dhurandhar et al., 2018] Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives.

[Doshi-Velez and Kim, 2017] Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science

of interpretable machine learning. arXiv preprint arXiv:1702.08608.

[Elliott et al., 2021] Elliott, A., Law, S., and Russell, C. (2021). Explaining classifiers using adversarial perturbations on the perceptual ball.

[Esser et al., 2020] Esser, P., Rombach, R., and Ommer, B. (2020). A disentangling invertible interpretation network for explaining latent representations.

[Evans et al., 2022] Evans, T., Retzlaff, C. O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.-R., Zerbe, N., and Holzinger, A. (2022). The explainability paradox: Challenges for xai in digital pathology. *Future Generation Computer Systems*, 133:281–296.

[Fong et al., 2019] Fong, R., Patrick, M., and Vedaldi, A. (2019). Understanding deep networks via extremal perturbations and smooth masks.

[Fong and Vedaldi, 2017] Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE.

[Geman et al., 2015] Geman, D., Geman, S., Hallonquist, N., and Younes, L. (2015). Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623.

[Gilpin, 2018] Gilpin, Leilani H. y Bau, D. y. Y. B. Z. y. B. A. y. S. M. y. K. L. (2018). Explicando explicaciones: Una visión general de la interpretabilidad del aprendizaje automático. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*.

[Goetschalckx et al., 2019] Goetschalckx, L., Andonian, A., Oliva, A., and Isola, P. (2019). Ganalyze: Toward visual definitions of cognitive image properties.

[Goodfellow et al., 2014a] Goodfellow, I., Shlens, J., and Szegedy, C. (2014a). Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

[Goodfellow et al., 2014b] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014b). Generative adversarial networks.

[Goyal et al., 2019] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Coun-

terfactual visual explanations.

[Guidotti, 2022] Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 38:1–55.

[He et al., 2018] He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. (2018). Attgan: Facial attribute editing by only changing what you want.

[Heusel et al., 2017] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*.

[Higaki et al., 2022] Higaki, A., Fukui, M., Tokuda, K., Mochizuki, T., Narita, R., Yamaguchi, S., Nakamori, S., Ishida, M., Kitagawa, K., Ichikawa, Y., and Sakuma, H. (2022). Myocardial perfusion imaging using artificial intelligence-generated synthetic images: A reader study. *Journal of Nuclear Cardiology*.

[Higgins et al., 2016] Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

[Horé and Ziou, 2010] Horé, A. and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In *2010 20th International Conference on Pattern Recognition*, pages 2366–2369.

[Hsieh et al., 2021] Hsieh, C.-Y., Yeh, C.-K., Liu, X., Ravikumar, P., Kim, S., Kumar, S., and Hsieh, C.-J. (2021). Evaluations and methods for explanation through robustness analysis.

[Huang et al., 2018] Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. (2018). Multimodal unsupervised image-to-image translation.

[Härkönen et al., 2020] Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. (2020). Ganspace: Discovering interpretable gan controls.

[Ignatov et al., 2024] Ignatov, A., Yates, J., and Boeva, V. (2024). Histopathological image classification with cell morphology aware deep neural networks. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 6913–6925.

[Isola et al., 2016] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976.

[Jahanian et al., 2020] Jahanian, A., Chai, L., and Isola, P. (2020). On the "steerability.ºf generative adversarial networks.

[Janowczyk and Madabhushi, 2016] Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of Pathology Informatics*, 7(1):29.

[Jha et al., 2020] Jha, A., Aicher, J. K., Gazzara, M. R., Singh, D., and Barash, Y. (2020). Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome Biology*, 21(1):149.

[Jolliffe and Cadima, 2016] Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202.

[Joshi et al., 2018] Joshi, S., Koyejo, O., Kim, B., and Ghosh, J. (2018). xgems: Generating examplars to explain black-box models.

[Kapishnikov et al., 2019] Kapishnikov, A., Bolukbasi, T., Viégas, F., and Terry, M. (2019). Xrai: Better attributions through regions.

[Karras et al., 2020a] Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020a). Training generative adversarial networks with limited data.

[Karras et al., 2019] Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks.

[Karras et al., 2020b] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020b). Analyzing and improving the image quality of stylegan.

[Kather et al., 2018] Kather, J. N., Halama, N., and Marx, A. (2018). 100,000 histological images of human colorectal cancer and healthy tissue.

[Khakzar et al., 2020] Khakzar, A., Baselizadeh, S., Khanduja, S., Rupprecht, C., Kim, S. T., and Navab, N. (2020). Improving feature attribution through input-specific network pruning.

[Kindermans et al., 2019] Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. (2019). *The (Un)reliability of Saliency Methods*, pages 267–280. Springer International Publishing, Cham.

[Komura and Ishikawa, 2018] Komura, D. and Ishikawa, S. (2018). Machine Learning Methods for Histopathological Image Analysis. *Computational and Structural Biotechnology Journal*, 16:34–42.

[Lample et al., 2018] Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., and Ranzato, M. (2018). Fader networks: Manipulating images by sliding attributes.

[Lang et al., 2021] Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W. T., Isola, P., Globerson, A., Irani, M., and Mosseri, I. (2021). Explaining in style: Training a gan to explain a classifier in stylespace.

[Lee et al., 2018] Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M. K., and Yang, M.-H. (2018). Diverse image-to-image translation via disentangled representations.

[Likert, 1932] Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.

[Lipton, 2018] Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3):31–57.

[Litjens et al., 2017] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

[Liu et al., 2019a] Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W., and Wen, S. (2019a). Stgan: A unified selective transfer network for arbitrary image attribute editing.

[Liu et al., 2018] Liu, M.-Y., Breuel, T., and Kautz, J. (2018). Unsupervised image-to-image

translation networks.

[Liu et al., 2019b] Liu, S., Kailkhura, B., Loveland, D., and Han, Y. (2019b). Generative counterfactual introspection for explainable deep learning.

[Liu et al., 2022] Liu, X., Sanchez, P., Thermos, S., O'Neil, A. Q., and Tsaftaris, S. A. (2022). Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80:102516.

[Lundervold and Lundervold, 2019] Lundervold, A. S. and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift für Medizinische Physik*, 29(2):102–127.

[Madabhushi and Lee, 2016] Madabhushi, A. and Lee, G. (2016). Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175. 20th anniversary of the Medical Image Analysis journal (MedIA).

[Madry et al., 2019] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks.

[McInnes et al., 2020] McInnes, L., Healy, J., and Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction.

[Metta et al., 2024] Metta, C., Beretta, A., Guidotti, R., Yin, Y., Gallinari, P., Rinzivillo, S., and Giannotti, F. (2024). Advancing dermatological diagnostics: Interpretable ai for enhanced skin lesion classification. *Diagnostics*, 14(7).

[Montavon et al., 2018] Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15.

[Mousavi et al., 2017] Mousavi, A., Dasarathy, G., and Baraniuk, R. G. (2017). Deepcodec: Adaptive sensing and recovery via deep convolutional neural networks.

[Muñoz et al., 2025a] Muñoz, B., Pezoa, R., and Gutierrez, H. (2025a). Histopathology image augmentation through stylegan2-ada. In Guerrero, G., San Martín, J., Meneses, E., Barrios Hernández, C. J., Osthoff, C., and Monsalve Diaz, J. M., editors, *High Performance*

Computing, pages 216–228, Cham. Springer Nature Switzerland.

[Muñoz et al., 2025b] Muñoz, B., Pezoa, R., and Gutierrez, H. (2025b). Histopathology image augmentation through stylegan2-ada. In Guerrero, G., San Martín, J., Meneses, E., Barrios Hernández, C. J., Osthoff, C., and Monsalve Diaz, J. M., editors, *High Performance Computing*, pages 216–228. Springer Nature Switzerland.

[Narayanaswamy et al., 2020] Narayanaswamy, A., Venugopalan, S., Webster, D. R., Peng, L., Corrado, G., Ruamviboonsuk, P., Bavishi, P., Sayres, R., Huang, A., Balasubramanian, S., Brenner, M., Nelson, P., and Varadarajan, A. V. (2020). Scientific discovery by generating counterfactuals using image translation.

[Nemirovsky et al., 2021] Nemirovsky, D., Thiebaut, N., Xu, Y., and Gupta, A. (2021). Countergan: Generating realistic counterfactuals with residual generative adversarial nets.

[Obermeyer et al., 2019] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

[Oh et al., 2021] Oh, K., Yoon, J. S., and Suk, H.-I. (2021). Born identity network: Multi-way counterfactual map generation to explain a classifier's decision.

[Pan et al., 2021] Pan, D., Li, X., and Zhu, D. (2021). Explaining deep neural network models with adversarial gradient integration. In *International Joint Conference on Artificial Intelligence*.

[Pawlowski et al., 2020] Pawlowski, N., Castro, D. C., and Glocker, B. (2020). Deep structural causal models for tractable counterfactual inference.

[Petsiuk et al., 2018] Petsiuk, V., Das, A., and Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models.

[Pichler and Hartig, 2022] Pichler, M. and Hartig, F. (2022). Machine learning and deep learning – a review for ecologists.

[Pidhorskyi et al., 2020] Pidhorskyi, S., Adjeroh, D., and Doretto, G. (2020). Adversarial latent

autoencoders.

[Plumerault et al., 2020] Plumerault, A., Borgne, H. L., and Hudelot, C. (2020). Controlling generative models with continuous factors of variations.

[PyTorch,] PyTorch. Pytorch pcam dataset.

[Qi et al., 2020] Qi, Z., Khorram, S., and Fuxin, L. (2020). Visualizing deep networks by optimizing with integrated gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11890–11898.

[Radford et al., 2016] Radford, A., Metz, L., and Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks.

[Ribeiro et al., 2016] Ribeiro, M., Singh, S., and Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In DeNero, J., Finlayson, M., and Reddy, S., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

[Rodriguez et al., 2021] Rodriguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I., Charlin, L., and Vazquez, D. (2021). Beyond trivial counterfactual explanations with diverse valuable explanations.

[Romero et al., 2019] Romero, A., Arbeláez, P., Gool, L. V., and Timofte, R. (2019). Smit: Stochastic multi-label image-to-image translation.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

[Samangouei et al., 2018] Samangouei, P., Saeedi, A., Nakagawa, L., and Silberman, N. (2018). Explaingan: Model explanation via decision boundary crossing transformations. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 681–696, Cham. Springer International Publishing.

[Schulz et al., 2020] Schulz, K., Sixt, L., Tombari, F., and Landgraf, T. (2020). Restricting the

flow: Information bottlenecks for attribution.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626.

[Seo et al., 2020] Seo, D., Oh, K., and Oh, I.-S. (2020). Regional multi-scale approach for visually pleasing explanations of deep neural networks. *IEEE Access*, 8:8572–8582.

[Shen et al., 2020] Shen, Z., Zhou, S. K., Chen, Y., Georgescu, B., Liu, X., and Huang, T. S. (2020). One-to-one mapping for unpaired image-to-image translation.

[Shitole et al., 2021] Shitole, V., Fuxin, L., Kahng, M., Tadepalli, P., and Fern, A. (2021). One explanation is not enough: Structured attention graphs for image classification.

[Shrikumar et al., 2019] Shrikumar, A., Greenside, P., and Kundaje, A. (2019). Learning important features through propagating activation differences.

[Shrikumar et al., 2017] Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2017). Not just a black box: Learning important features through propagating activation differences.

[Simonyan et al., 2014] Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps.

[Singla et al., 2020] Singla, S., Pollack, B., Chen, J., and Batmanghelich, K. (2020). Explanation by progressive exaggeration.

[Sloboda et al., 2024] Sloboda, T., Hudec, L., Halinkovič, M., and Benesova, W. (2024). Attention-enhanced unpaired xai-gans for transformation of histological stain images. *Journal of Imaging*, 10(2).

[Smilkov et al., 2017] Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise.

[Springenberg et al., 2015] Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2015). Striving for simplicity: The all convolutional net.

[Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks.

[Tanyel et al., 2023] Tanyel, T., Ayvaz, S., and Keserci, B. (2023). Beyond known reality: Exploiting counterfactual explanations for medical research.

[Tellez et al., 2019] Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., and van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544.

[TensorFlow,] TensorFlow. Tensorflow datasets: Pcam.

[Thakur and Fischmeister, 2021] Thakur, S. and Fischmeister, S. (2021). A generalizable saliency map-based interpretation of model outcome. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 4099–4106.

[Tizhoosh and Pantanowitz, 2018] Tizhoosh, H. R. and Pantanowitz, L. (2018). Artificial intelligence and digital pathology: Challenges and opportunities. *Journal of Pathology Informatics*, 9(1):38.

[Tov et al., 2021] Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., and Cohen-Or, D. (2021). Designing an encoder for stylegan image manipulation.

[Turing, 2009] Turing, A. M. (2009). *Computing Machinery and Intelligence*, pages 23–65. Springer Netherlands, Dordrecht.

[Verma et al., 2021] Verma, S., Dickerson, J., and Hines, K. (2021). Counterfactual explanations for machine learning: Challenges revisited.

[Voynov and Babenko, 2020] Voynov, A. and Babenko, A. (2020). Unsupervised discovery of interpretable directions in the gan latent space.

[Wachter et al., 2017] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Cybersecurity*.

[Wang and Vasconcelos, 2020] Wang, P. and Vasconcelos, N. (2020). Scout: Self-aware discri-

minant counterfactual explanations.

[Wang et al., 2004] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

[Wei et al., 2019] Wei, J., Suriawinata, A., Vaickus, L., Ren, B., Liu, X., Wei, J., and Hassanpour, S. (2019). Generative image translation for data augmentation in colorectal histopathology images.

[Wei et al., 2018] Wei, Y., Chang, M.-C., Ying, Y., Lim, S. N., and Lyu, S. (2018). Explain black-box image classifications using superpixel-based interpretation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1640–1645.

[Woods et al., 2019] Woods, W., Chen, J., and Teuscher, C. (2019). Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nature Machine Intelligence*, 1(11):508–516.

[Wu et al., 2020] Wu, Z., Lischinski, D., and Shechtman, E. (2020). Stylespace analysis: Disentangled controls for stylegan image generation.

[Xiao et al., 2018] Xiao, T., Hong, J., and Ma, J. (2018). Elegant: Exchanging latent encodings with gan for transferring multiple face attributes.

[Xu et al., 2020] Xu, S., Venugopalan, S., and Sundararajan, M. (2020). Attribution in scale and space.

[Xue et al., 2021] Xue, Y., Ye, J., Zhou, Q., Long, L. R., Antani, S., Xue, Z., Cornwell, C., Zaino, R., Cheng, K. C., and Huang, X. (2021). Selective synthetic augmentation with histogan for improved histopathology image classification. *Medical Image Analysis*, 67:101816.

[Yang et al., 2021] Yang, F., Liu, N., Du, M., and Hu, X. (2021). Generative counterfactuals for neural networks via attribute-informed perturbation.

[Yang et al., 2020] Yang, Q., Zhu, X., Fwu, J.-K., Ye, Y., You, G., and Zhu, Y. (2020). Mfpp: Morphological fragmental perturbation pyramid for black-box model explanations.

[Yao et al., 2021] Yao, X., Newson, A., Gousseau, Y., and Hellier, P. (2021). A latent transformer for disentangled face editing in images and videos.

[Yu et al., 2018] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. (2018). Generative image inpainting with contextual attention.

[Yu et al., 2019] Yu, X., Chen, Y., Li, T., Liu, S., and Li, G. (2019). Multi-mapping image-to-image translation via learning disentanglement.

[Yu et al., 2020] Yu, X., Ying, Z., Li, T., Liu, S., and Li, G. (2020). Multi-mapping image-to-image translation with central biasing normalization.

[Zeiler and Fergus, 2013] Zeiler, M. D. and Fergus, R. (2013). Visualizing and understanding convolutional networks. *ArXiv*, abs/1311.2901.

[Zhang et al., 2018] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric.

[Zhang et al., 2021] Zhang, Y., Khakzar, A., Li, Y., Farshad, A., Kim, S. T., and Navab, N. (2021). Fine-grained neural network explanation by identifying input features with predictive information.

[Zhou et al., 2015] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Object detectors emerge in deep scene cnns.

[Zhou et al., 2016] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929.

[Zhu et al., 2020] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2020). Unpaired image-to-image translation using cycle-consistent adversarial networks.

[Zintgraf et al., 2017] Zintgraf, L. M., Cohen, T., Adel, T., and Welling, M. (2017). Visualizing deep neural network decisions: Prediction difference analysis. *ArXiv*, abs/1702.04595.

Apéndices

Apéndice A

A.1. Métodos de Explicaciones Visuales como Atribución de Características

A.1.1. Saliency Methods

Los *saliency methods* utilizan la retropropagación basada en gradientes para identificar cómo cada píxel i de una imagen de entrada x afecta la salida de un modelo entrenado f. La idea principal es calcular los gradientes de la predicción con respecto a una imagen de entrada y así determinar la magnitud del impacto de cada píxel en la salida de clasificación $f(x)$. Estos gradientes se utilizan para generar un mapa que muestra la relevancia de cada píxel.

Gradient-baseline: Uno de los métodos pioneros se presenta en [Simonyan et al., 2014], conocido como Vanilla Gradient. Este calcula directamente las derivadas de puntaje generado por el modelo para una clase específica con respecto a cada píxel de la imagen de entrada. La atribución de cada píxel i se calcula como:

$$\mathcal{E}_i(x) = \frac{\partial f_c}{\partial x_i}(x)$$

donde,

- $\mathcal{E}_i(x)$: Representa la contribución del píxel i a la salida de clasificación $f_c(x)$.
- $f_c(x)$: Es la predicción de clasificación del modelo f para a la clase c.
- x_i: Es el valor de la imagen x en el píxel i.

Sin embargo, debido a la naturaleza no lineal de las redes neuronales, este método enfrenta problemas como el del gradiente desvaneciente (vanishing gradient) y puntos singulares, lo cual lleva a explicaciones ruidosas y difíciles de interpretar. Existen varias contribuciones que ponen foco en construir mapas de explicabilidad más suaves y precisos (ver Figura A.1).

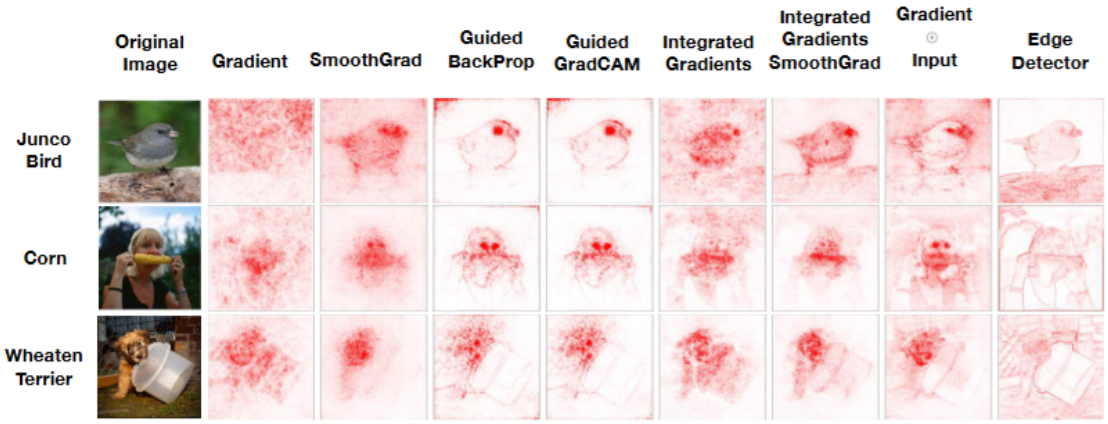


Figura A.1: **Comparación de los saliency methods más comunes y un edge detector [Adebayo et al., 2020].** Saliency masks aplicados a tres entradas diferentes para un modelo Inception v3 entrenado en ImageNet.

En [Bach et al., 2015] se propone Layer-wise Relevance Propagation (LRP), método que retropropaga puntuaciones de relevancia en vez del gradiente, es decir, LRP nos dice específicamente cuánto contribuye cada píxel en la predicción en vez de describir el impacto de los cambios en cada píxel. El funcionamiento de LRP se basa en un principio de conservación similar a la ley de conservación de la energía en física. Se inicia con la predicción final de la red y se propaga hacia atrás a través de las capas hasta llegar a la entrada, preservando la suma total en cada paso. La retropropagación sigue un conjunto de reglas específicas adaptadas en cada capa, diversos trabajos proponen diversas reglas [Montavon et al., 2018].

En [Shrikumar et al., 2017] se presenta DeepLIFT, un método alternativo a LRP que propaga una señal de importancia a través de la red desde la salida hasta la entrada, calculando las diferencias entre la entrada y una imagen de referencia neutra, por ejemplo, una imagen negra. Esta configuración ayuda al método a evitar artefactos relacionados con discontinuidades en el gradiente. En la práctica, se asigna una atribución a cada neurona en todas las capas (como en LRP), la cual refleja la diferencia de activación en la red neuronal, entre los valores de referencia y los valores de la entrada. Estos valores de referencia para todas la unidades ocultas se obtienen alimentando la red con la entrada de referencia.

Path-based methods: Diversos path-based methods han surgido como extensiones y mejoras del enfoque de Gradientes Integrados propuesto por [Sundararajan et al., 2017]. Este método aborda los problemas del desvanecimiento del gradiente. Primero, se toma una ruta recta entre dos puntos, un punto de referencia inicial $\bar{x}$ (llamado línea base) y la imagen x que queremos analizar. Luego, se calcula cómo contribuye cada característica de la entrada a lo largo de esta ruta recta. La ruta se define matemáticamente como:

$$\gamma(\lambda) = \bar{x} + \lambda(x - \bar{x}),$$

donde λ va cambiando gradualmente de 0 a 1 para movernos desde el punto base hasta la imagen final. Una vez que se promedian todas las distribuciones, la explicación visual en el píxel i se expresa como:

$$\mathcal{E}_i(x) = \int_0^1 \frac{\partial f_c(\gamma(u))}{\partial \gamma(u)} \frac{\partial \gamma(u)}{\partial u}\, du = (x_i - \bar{x}_i) \int_0^1 \frac{\partial f_c}{\partial x_i}(\gamma(u))\, du$$

Luego, en [Jha et al., 2020] los autores emplean un autoencoder variacional para proyectar las imágenes en un espacio latente. En este espacio, definen distintas estrategias para la línea base del Gradiente Integrado, como los puntos cero, la mediana entre clases o el promedio de los k vecinos más cercanos. Posteriormente, transforman esta línea base al espacio de entrada utilizando el decodificador del autoencoder, lo que permite una referencia más informativa y adaptada a la estructura de los datos.

En [Kapishnikov et al., 2019], los autores calculan los gradientes integrados para dos líneas ba-

se: una imagen completamente negra y otra completamente blanca. Argumentan que todos los píxeles deben tener la misma probabilidad de contribuir, independientemente de la línea base seleccionada. Esto evita que la proximidad de los píxeles a una línea base específica introduzca sesgos. Además, integran este enfoque con un algoritmo de segmentación, donde generan múltiples segmentos de la entrada. Inician con una máscara vacía y, de manera iterativa, agregan la región del segmento que maximiza la atribución total. Este procedimiento da lugar a mapas de explicación más suaves y coherentes.

En [Xu et al., 2020], se propone un método que genera atribuciones tanto en frecuencia como en espacio. Esto permite identificar características relevantes en niveles de granularidad gruesos y finos. Su enfoque considera una ruta que conecta la imagen de entrada con una versión borrosa obtenida mediante un filtro Gaussiano. La línea base es la imagen más borrosa, que carece de información significativa, lo que asegura que la atribución esté bien fundamentada.

En [Pan et al., 2021], los autores replantean los Gradientes Integrados enfocándose en identificar qué distingue la clase predicha del resto, en lugar de explicar qué conduce al modelo a elegir dicha clase. Para ello, demuestran que el gradiente asociado con la clase predicha equivale a la suma negativa de los gradientes de todas las clases adversas. Este enfoque utiliza rutas no lineales, siguiendo técnicas similares a las empleadas para generar ataques adversarios [Madry et al., 2019]. Las rutas se definen mediante pasos de gradiente con dirección signada hacia las clases adversas, lo que permite una integración más rica de información contextual y explicativa.

En [Smilkov et al., 2017] se propone el método SmoothGrad. Esta es una técnica que mejora los saliency maps reduciendo el ruido mediante la adición de perturbaciones Gaussianas a la entrada. Se generan múltiples copias del input añadiendo ruido $\epsilon_k \sim N(0, \sigma^2)$, y para cada una se calcula el saliency map. Luego, se promedian los resultados, obteniendo un mapa más suave y coherente:

$$\mathcal{E}_i(x) = \frac{1}{n} \sum_{k=1}^{n} \frac{\partial f_u}{\partial x_i} (x + \epsilon_k).$$

De esta manera, se obtiene una visualización más estable y clara al reducir la influencia de

pequeñas variaciones locales en la imagen que podrían generar gradientes irrelevantes. Aunque SmoothGrad logra mejorar la calidad visual de los saliency maps, su principal desventaja es el costo computacional adicional, ya que requiere múltiples ejecuciones del modelo para obtener un promedio significativo.

Existen múltiples enfoques alternativos para generar saliency maps, cada uno con distintas estrategias. En [Zeiler and Fergus, 2013] se propone un método que utiliza deconvoluciones para visualizar las activaciones de capas específicas y su relación con la entrada. Las operaciones de pooling se invierten para recuperar información espacial y generar mapas de saliencia interpretables. Luego, en [Springenberg et al., 2015] se introduce una variante de retropropagación que filtra gradientes negativos, asegurando que solo las señales positivas de activación y gradiente contribuyan al saliency map. En [Bojarski et al., 2018], los autores generan mapas de saliencia promediando las activaciones ReLU en todas las capas del modelo. Utilizan deconvolución para proyectar estas activaciones hacia el espacio de entrada de manera iterativa, multiplicando el mapa de características promediado de cada capa con el mapa sobremuestreado de la capa anterior. Este proceso combina información jerárquica de múltiples niveles, destacando las regiones de entrada que activan consistentemente características importantes.

Los métodos basados en gradientes, como los propuestos por [Simonyan et al., 2014, Sundararajan et al., 2017, Smilkov et al., 2017], requieren únicamente acceso a los gradientes del modelo, los cuales se retropropagan a través de la red neuronal para calcular las atribuciones. En contraste, otras técnicas demandan un conocimiento más profundo de la arquitectura del modelo. Por ejemplo, algunos métodos necesitan acceso completo para calcular gradientes discretos específicos, como en [Bach et al., 2015, Shrikumar et al., 2017], mientras que otros construyen redes deconvolucionales adaptadas, como se observa en [Zeiler and Fergus, 2013, Springenberg et al., 2015, Bojarski et al., 2018]. Además, [Kindermans et al., 2019] evidencian que ciertas transformaciones en los datos de entrada, aunque no afecten directamente al modelo, como añadir un desplazamiento constante, pueden alterar significativamente los resultados de algunos métodos de atribución.

Tabla A.1: Comparación global de los saliency methods. Para cada método, se indica si la explicación visual es local o global y si el método es agnóstico al modelo (se muestra a qué estructuras internas se tiene acceso). Notar que NN se refiere a una red neuronal.

Método	**Tipo Expl.**	**Model Agnostic**
Deconvnet. [Zeiler and Fergus, 2013]	Local	Acceso a NN
Gradient baseline [Simonyan et al., 2014]	Local	Acceso a Gradientes
LRP [Bach et al., 2015]	Local	Acceso a Gradientes + NN
Guided Backprop. [Springenberg et al., 2015]	Local	Acceso a Gradientes + NN
Input x Grad. [Shrikumar et al., 2017]	Local	Acceso a Gradientes
DeepLIFT [Shrikumar et al., 2019]	Local	Acceso a Gradientes + NN
Smooth Grad. [Smilkov et al., 2017]	Local	Acceso a Gradientes
IG [Sundararajan et al., 2017]	Local	Acceso a Gradientes
Visual Backprop. [Bojarski et al., 2018]	Local	Acceso a NN
XRAI [Kapishnikov et al., 2019]	Local	Acceso a Gradientes
Enhanced IG [Jha et al., 2020]	Local/Global	Acceso a Gradientes
Blur IG [Xu et al., 2020]	Local	Acceso a Gradientes
AIG [Pan et al., 2021]	Local	Acceso a Gradientes

A.1.2. Class Activation Map Methods

Los métodos de class activation mapping (CAM) o mapas de activación de clase (en español) son un enfoque destacado en las explicaciones visuales para redes neuronales convolucionales (CNNs). Estos métodos tienen como objetivo localizar las regiones de una imagen que son más relevantes para la predicción del modelo. Al analizar las activaciones de la última capa convolucional, los métodos basados en CAM pueden generar visualizaciones que destacan las áreas de la imagen que más contribuyen a una decisión de clasificación en particular. Las unidades de las capas convolucionales en Redes Neuronales Convolucionales (CNN) de clasificación pueden

capturar información sobre la localización de objetos dentro de las imágenes [Zhou et al., 2015]. Sin embargo, esta información suele perderse en las capas finales completamente conectadas, donde se genera la salida de clasificación.

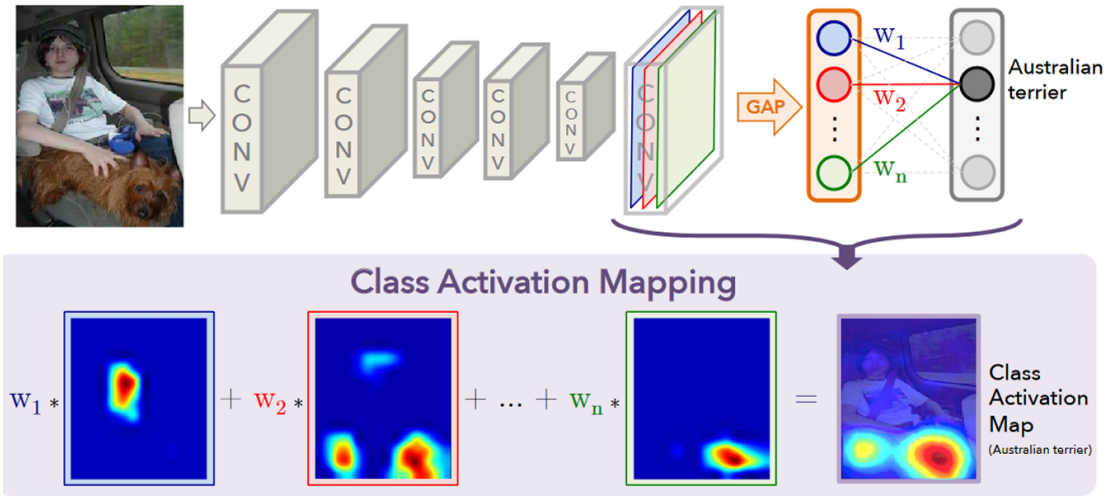


Figura A.2: **Vista General del Método Class Activation Mapping [Zhou et al., 2016].** Para una red neuronal que utiliza una operación de Global Average Pooling (GAP) para calcular las salidas del modelo, la puntuación de la clase predicha se proyecta hacia atrás y se combina con la última capa convolucional para producir Class Activation Maps (CAM) o mapas de activación de clase, en español. Luego, este mapa de baja resolución se reescala al tamaño de la entrada para generar el mapa de atribución.

Para preservar dicha información, [Zhou et al., 2016] proponen reemplazar las capas completamente conectadas por una operación de global average pooling (GAP) como capa final de la red. Este enfoque permite calcular un mapa de activación de clase para una categoría específica (como la clase predicha) directamente desde la última capa convolucional, manteniendo la relación entre las características espaciales y su relevancia para la predicción. El mapa de activación de clase se amplía mediante sobremuestreo hasta coincidir con las dimensiones de la imagen de entrada, produciendo un mapa de explicación. Este mapa facilita identificar visualmente las regiones más relevantes de la imagen que contribuyen a la predicción de la clase específica,

mejorando así la interpretabilidad del modelo.

Para comprender la importancia de las neuronas en una red convolucional, GradCAM [Selvaraju et al., 2017] extiende el trabajo presentado en [Zhou et al., 2016] calculando el gradiente de la salida del modelo para una clase específica con respecto a la última capa convolucional. La elección de esta capa se justifica porque equilibra adecuadamente la preservación de la información espacial y la representación semántica de alto nivel. GradCAM generaliza el enfoque de CAM y es aplicable a cualquier red convolucional, destacando las regiones importantes de la imagen para la decisión de clasificación. No obstante, este método presenta limitaciones. Por ejemplo, a menudo no logra localizar múltiples ocurrencias de una misma clase dentro de la misma imagen ni capturar el objeto completo, generando mapas gruesos que carecen de detalles de grano fino. GradCAM++, introducido en [Chattopadhay et al., 2018], aborda estas deficiencias al incluir una ponderación de los gradientes a nivel de píxel, lo que permite una localización más refinada de las características relevantes. A pesar de estas mejoras, la resolución de los mapas generados sigue siendo limitada, lo que afecta su utilidad en casos que requieren mayor precisión.

Los enfoques mencionados producen explicaciones visuales al sobremuestrear el mapa de activación de clase desde la última capa convolucional hasta que coincide con el tamaño de la entrada. Para refinar los resultados, [Selvaraju et al., 2017] y [Chattopadhay et al., 2018] combinan sus métodos con la retropropagación guiada (Guided Backpropagation) [Springenberg et al., 2015], logrando mayor claridad en las regiones destacadas. Sin embargo, CAM tiene la limitación de depender de una arquitectura específica y no ser agnóstico al modelo, mientras que los otros métodos de mapas de activación de clase requieren acceso a gradientes y capas profundas, lo que restringe su aplicabilidad universal.

Tabla A.2: Comparación global de métodos CAM. Para cada método, indicamos si la explicación visual es local o global, si el usuario tiene acceso a las estructuras internas del modelo o no, y si necesitamos datos adicionales o solo los datos probados.

Método	**Tipo Expl.**	**Model Agnostic**
CAM [Zhou et al., 2016]	Local	Específico para CNN
GradCAM [Selvaraju et al., 2017]	Local	Acceso a Gradientes + CNN
GradCAM++ [Chattopadhay et al., 2018]	Local	Acceso a Gradientes + CNN
Guided GradCAM [Selvaraju et al., 2017]	Local	Acceso a Gradientes + CNN

A.1.3. Métodos Basados en Perturbaciones

Los métodos basados en perturbaciones (perturbation methods) son un enfoque fundamental dentro de las explicaciones visuales en redes neuronales, enfocados en modificar la entrada para observar cómo cambia la salida del modelo. A través de estos métodos, se busca identificar qué partes de la imagen son esenciales para la predicción, proporcionando así una medida directa de la importancia de cada píxel o región. A diferencia de los métodos basados en gradientes, los métodos de perturbación son generalmente más intuitivos, ya que se basan en la eliminación o alteración de partes de la imagen para medir su impacto en la predicción. Esta sección explora los principales métodos de perturbación utilizados para generar explicaciones visuales, basándonos en una selección de trabajos destacados en la literatura.

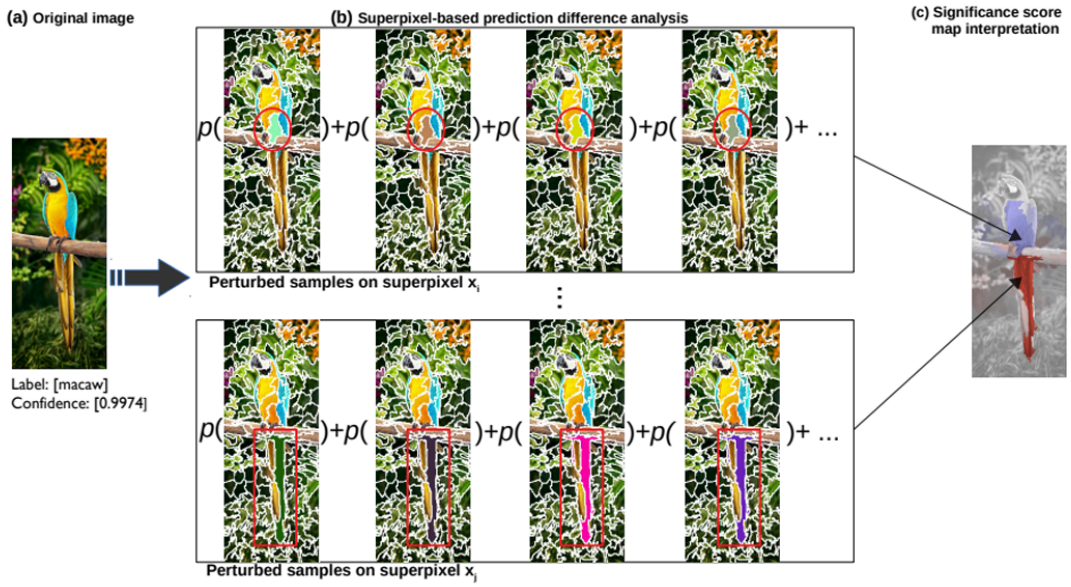


Figura A.3: **Explicación mediante el enfoque de perturbación por superpíxeles [Wei et al., 2018].** (a) La imagen de entrada y la predicción de clasificación correspondiente. (b) Se calcula el análisis de diferencia en la predicción para cada superpíxel de entrada (la figura ilustra dos de ellos). La puntuación de relevancia de cada segmento es el promedio de distintas operaciones de marginalización (ausencia de características). (c) El mapa de explicación resultante muestra las regiones que apoyan o no apoyan la decisión del modelo.

Estos enfoques se fundamentan en el análisis de diferencia de predicción (PDA, por sus siglas en inglés) y típicamente generan una máscara óptima M, que indica las áreas donde una función de perturbación Φ debería actuar para alterar la predicción del modelo. Esta máscara M sirve como mapa de explicación visual. A continuación, se resumen las principales contribuciones en este ámbito.

Una técnica pionera en este campo es la **oclusión** [Zeiler and Fergus, 2013], que consiste en ocultar/eliminar sistemáticamente pequeños parches de la imagen de entrada reemplazando los píxeles por valores constantes como negro, y medir los cambios en la predicción. Si la oclusión de una determinada región provoca una caída significativa en la predicción del modelo, se puede

inferir que esa región es crucial para la predicción. De manera similar en [Zintgraf et al., 2017] los autores evalúan la evidencia de clase a través de parches, aunque en este caso, los valores de los píxeles dentro de cada parche se muestrean de su vecindad local, aprovechando la dependencia espacial entre píxeles.

A pesar de su efectividad, estos métodos tienen un alto costo computacional, ya que la operación de perturbación debe repetirse para cada píxel o parche a lo largo de toda la entrada. Además, el rendimiento y los resultados dependen del tamaño y la superposición de los parches elegidos. Varias extensiones de estas técnicas buscan mejorar su eficiencia al utilizar diferentes regiones o tipos de perturbaciones.

En [Wei et al., 2018], se propone un enfoque alternativo que utiliza regiones de superpíxeles en lugar de parches. La entrada se segmenta en áreas coherentes con el contenido de la imagen. Para cada superpíxel, calculan una predicción promedio. Luego, asignan valores RGB aleatorios al superpíxel, muestreados del histograma de entrada, y evalúan el modelo. Este enfoque reduce la granularidad y el costo computacional en comparación con la oclusión de píxeles individuales.

En [Seo et al., 2020], los autores amplían la segmentación al usar múltiples escalas de superpíxeles, dividiendo la entrada en secciones que van desde los 2 a los 2^r segmentos (con $r = 5$). Cada segmento se analiza reemplazando los valores de los superpíxeles con muestras de una distribución normal que aproxima los valores originales. Las regiones seleccionadas se sobremuestrean al tamaño de la entrada, y se buscan combinaciones de regiones que preserven la salida del modelo. Las regiones que conservan la predicción intacta se dejan sin modificar, mientras que el resto de la imagen se rellena con píxeles negros. El borde de las máscaras sobremuestreadas toma valores en $[0, 1]$ para suavizar la transición. Además, las regiones seleccionadas se analizan mediante grafos de atención estructurados (Structured Attention Graphs). Esto se logra eliminando iterativamente subregiones dentro de las áreas preservadas y evaluando cómo estas eliminaciones afectan la predicción del modelo.

Los métodos de muestreo de regiones múltiples generan explicaciones visuales al muestrear diversas regiones de la imagen de entrada, medir el impacto de su perturbación y promediar sus contribuciones. Este enfoque permite identificar la importancia relativa de diferentes áreas de la

imagen para la predicción del modelo.

Uno de ellos es LIME (Local Interpretable Model-agnostic Explanations) [Ribeiro et al., 2016], que introduce un enfoque en el que se perturban superpíxeles aleatorios de la imagen de entrada. A partir de estas perturbaciones, se entrena un clasificador lineal local que estima la importancia de cada segmento para la predicción del modelo. Este método ofrece explicaciones locales agnósticas al modelo, capturando la relevancia de las diferentes regiones para la predicción de una clase específica.

RISE (Randomized Input Sampling for Explanation) [Petsiuk et al., 2018] propone una técnica en la que se generan múltiples máscaras M_i, que luego se aplican a la entrada x mediante una operación de producto por elementos ($x \odot M_i$). Para cada máscara, se calcula la salida del modelo f hacia una clase específica c. Luego, el mapa de explicación $E(x)$ se define como:

$$E(x) = \frac{1}{E[M] \cdot N} \sum_{i=1}^{N} f_c(x \odot M_i) \cdot M_i$$

Donde,

- $\odot$ denota la multiplicación por elementos.
- M_i se genera como una máscara binaria de baja resolución que posteriormente se sobremuestrea para coincidir con el tamaño de la entrada.
- $f_c(x \odot M_i)$ representa la predicción del modelo para la clase c sobre la entrada enmascarada.

En la misma línea, MFPP (Multi-scale Feature Perturbation Process) [Yang et al., 2020] amplía la idea presentada en RISE al tener en cuenta la estructura de la imagen en la generación de las máscaras. En lugar de crear regiones de máscara aleatorias, este método utiliza un algoritmo de segmentación para producir máscaras de superpíxeles a diferentes escalas. El proceso consiste en generar múltiples máscaras basadas en superíxeles segmentados para luego calcular las predicciones del modelo f_c para cada entrada enmascarada y finalmente obtener el mapa de explicación final promediando la contribución de todas las máscaras en todas las escalas. Al incorporar la estructura de la imagen en el proceso de generación de máscaras, MFPP mejora la

calidad y precisión de las explicaciones en comparación con RISE, que ignora esta información contextual.

A.1.4. Métodos de Ejemplos Adversariales

Los ejemplos adversariales (Adversarial Examples) es un método que en lugar de optimizar una máscara a través de una función de perturbación, se basa en encontrar un ejemplo adversario cercano para cada imagen de entrada. Este ejemplo adversario se compara con la entrada en términos de la distancia L_p, donde p puede ser [1, 2, ..., ∞], y está diseñado para impactar la decisión del clasificador dentro de un espacio restringido.

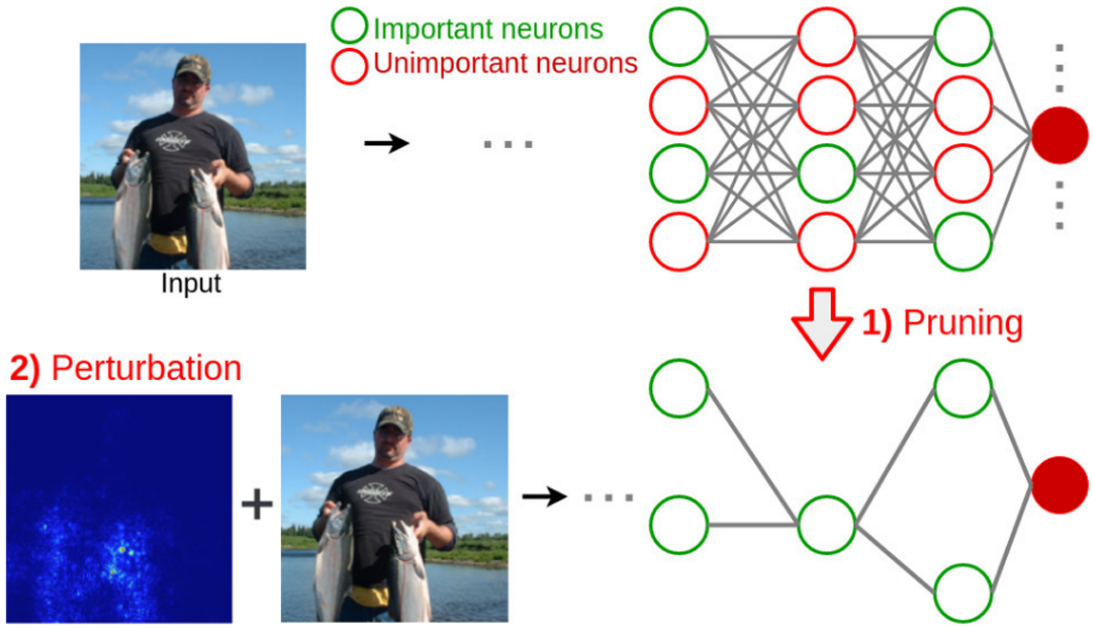


Figura A.4: **Perturbación de las características importantes mediante una estrategia de poda [Khakzar et al., 2020].** (1) Eliminación de las neuronas no importantes para la predicción desde el modelo de clasificación. (2) Generación de una perturbación adversaria que cambia al máximo la salida de la red podada.

Estos enfoques aprovechan técnicas de generación de ejemplos adversarios, como las propues-

tas en [Goodfellow et al., 2014a] y [Madry et al., 2019], cuyo objetivo principal es generar una perturbación mínima δ (casi siempre imperceptible) que, al sumarse a la entrada, pueda engañar al modelo. El problema de optimización asociado se formula de la siguiente manera:

$$\delta^* = \arg\max_{\delta \epsilon A} L_f(x + \delta, c)$$

Aquí:

- A es el espacio de ataques permitido.
- L_f es un término de clasificación relacionado con el modelo de clasificación f.
- c es la clase predicha de la entrada x (o la clase verdadera).

En ataques adversarios tradicionales, la perturbación δ se busca minimizando su norma L_p. Sin embargo, estas perturbaciones no están optimizadas para destacar características relevantes de la entrada para el modelo f.

En [Woods et al., 2019], se introduce un enfoque que optimiza la forma de la perturbación siguiendo el gradiente $\partial f_c / \partial(x + \delta)$, respetando un límite de RMSE. Este límite asegura que la perturbación sea mínima mientras permite una cierta cantidad de diferencia perceptual con la entrada.

En [Elliott et al., 2021], los autores proponen el uso de perturbaciones perceptuales para capturar patrones relevantes en la entrada, reduciendo al mismo tiempo los artefactos adversarios. Esto se logra mediante la adición de una pérdida perceptual como regularización adicional. Para destacar las regiones más relevantes en las explicaciones adversarias, se aplica un desenfoque Gaussiano con parámetros σ. Esto resalta las diferencias más importantes entre x y x_a, permitiendo obtener una explicación visual significativa.

En [Khakzar et al., 2020], los autores proponen un método en el que primero podan las neuronas no esenciales del modelo, es decir, aquellas que tienen un impacto mínimo en la salida del modelo. Luego, optimizan una perturbación de entrada que altera al máximo la salida del modelo podado. En este escenario, solo las neuronas importantes se ven afectadas por el ataque adversario.

A través de un análisis de robustez adversaria, en [Hsieh et al., 2021] se propone una técnica para evaluar las atribuciones de las explicaciones. Plantean dos suposiciones clave: La primera es que cuando los valores de las características relevantes se mantienen fijos, las perturbaciones aplicadas al resto de las características tienen poco impacto en la salida del modelo. La segunda, por el contrario, es que incluso pequeñas perturbaciones aplicadas a las características relevantes (mientras se mantienen fijas las demás) deberían afectar significativamente la salida del modelo. Luego, derivan explicaciones visuales que maximizan los criterios de evaluación utilizando un algoritmo iterativo para agregar elementos al conjunto de características objetivo, un proceso que es costoso computacionalmente.

En comparación con los ataques adversarios convencionales, estos enfoques introducen cierta regularización, incentivando que las perturbaciones se apliquen solo a las características esenciales para el modelo. Sin embargo, su formulación no garantiza que los ejemplos adversarios generados sean consistentes con las distribuciones reales de los datos. Para abordar esto, imponen explícitamente una restricción de distancia por píxel (distancias L_p) entre el ejemplo adversario y la imagen original, es decir, $\|x - x_{\text{adv}}(x)\|_p$, donde $p \in \{1, 2\}$. A pesar de la regularización, esto restringe en exceso la generación de ejemplos adversarios y a menudo produce artefactos. Además, dificulta la captura de patrones específicos de la distribución.

Tabla A.3: Comparación global de métodos de Perturbación. Para cada método, se indica si la explicación visual es local o global y si el usuario tiene acceso a las estructuras internas del modelo (y cuáles). "NN as Encoder"significa que la red neuronal estudiada se usa como parte encoder de un modelo generativo.

Método	**Tipo Expl.**	**Model Agnostic**
Occlusion [Zeiler and Fergus, 2013]	Local	✓
Pixel PDA [Zintgraf et al., 2017]	Local	✓
Superpixel PDA [Wei et al., 2018]	Local	✓
Multiscale PDA [Seo et al., 2020]	Local	✓
SAG [Shitole et al., 2021]	Local	✓
LIME [Ribeiro et al., 2016]	Local	✓
RISE [Petsiuk et al., 2018]	Local	✓
MFPP [Yang et al., 2021]	Local	✓
ERM Mask [Thakur and Fischmeister, 2021]	Local	✓
BBMP [Fong and Vedaldi, 2017]	Local	✓
Extr. Pert. [Fong et al., 2019]	Local	NN access
I-GOS [Qi et al., 2020]	Local	✓
Mask Generator [Dabkowski and Gal, 2017]	Local	✓
IBA, inverse IBA [Schulz et al., 2020]	Local/Global	NN access
Fine-grained IBA [Zhang et al., 2021]	Local	NN access

Apéndice B

B.1. Tabla de Tiempos de Entrenamiento Según Resolución y Cantidad de GPUs

Resolution	GPUs	1000 kimg	25000 kimg	sec/kimg	GPU mem	CPU mem
128×128	1	4h 05m	4d 06h	12.8–13.7	7.2 GB	3.9 GB
	2	2h 06m	2d 04h	6.5–6.8	7.4 GB	7.9 GB
	4	1h 20m	1d 09h	4.1–4.6	4.2 GB	16.3 GB
256×256	1	6h 36m	6d 21h	21.6–24.2	5.0 GB	4.5 GB
	2	3h 27m	3d 14h	11.2–11.8	5.2 GB	9.0 GB
	4	1h 45m	1d 20h	5.6–5.9	5.2 GB	17.8 GB
512×512	1	21h 03m	21d 22h	72.5–74.9	7.6 GB	5.0 GB
	2	10h 59m	11d 10h	37.7–40.0	7.8 GB	9.8 GB
	4	5h 29m	5d 17h	18.7–19.1	7.9 GB	17.7 GB
1024×1024	1	1d 20h	46d 03h	154.3–161.6	8.1 GB	5.3 GB
	2	23h 09m	24d 02h	80.6–86.2	8.6 GB	11.9 GB
	4	11h 36m	12d 02h	40.1–40.8	8.4 GB	21.9 GB

Tabla B.1: Configuraciones y recursos de entrenamiento con StyleGAN2-ADA [Karras et al., 2020b] utilizando distintos números de GPUs.

Apéndice C

C.1. Software de evaluación con expertos: Generación y evaluación de Imágenes Contrafactuales

A continuación se mostrarán diferentes pantallazos de cada sección del software de evaluación con expertos.

Figura C.1: Pantallazo de las instrucciones iniciales del software.

Figura C.2: Pantallazo de las instrucciones de la sección 1 del software.

Figura C.3: Pantallazo de la primera evaluación de la sección 1 del software.

Figura C.4: Pantallazo de las instrucciones de la sección 2 del software.

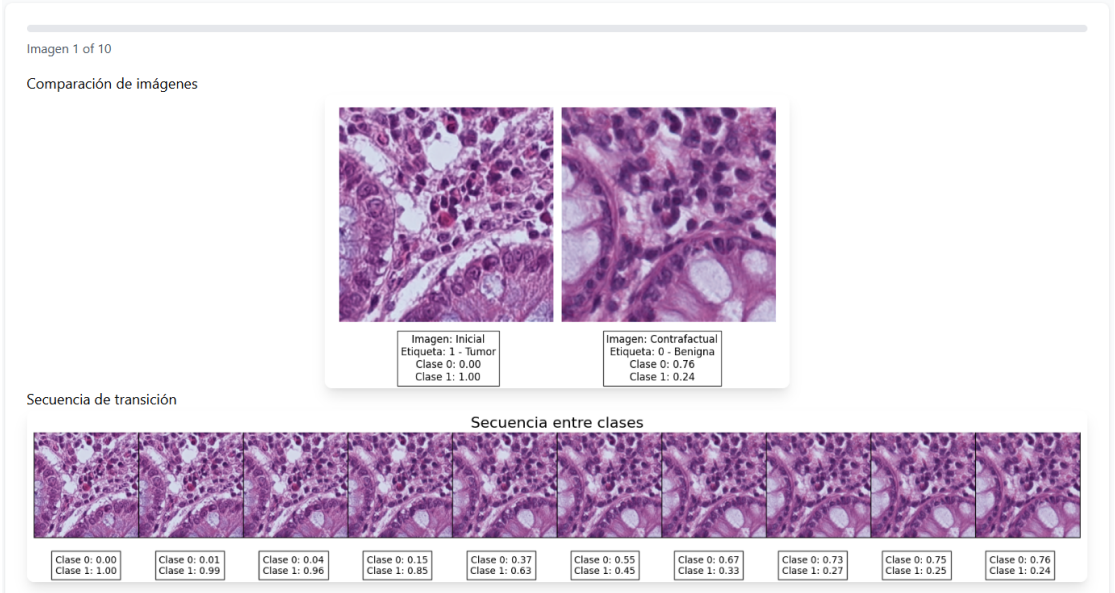


Figura C.5: Pantallazo de la primera evaluación de la sección 2 del software, parte 1.

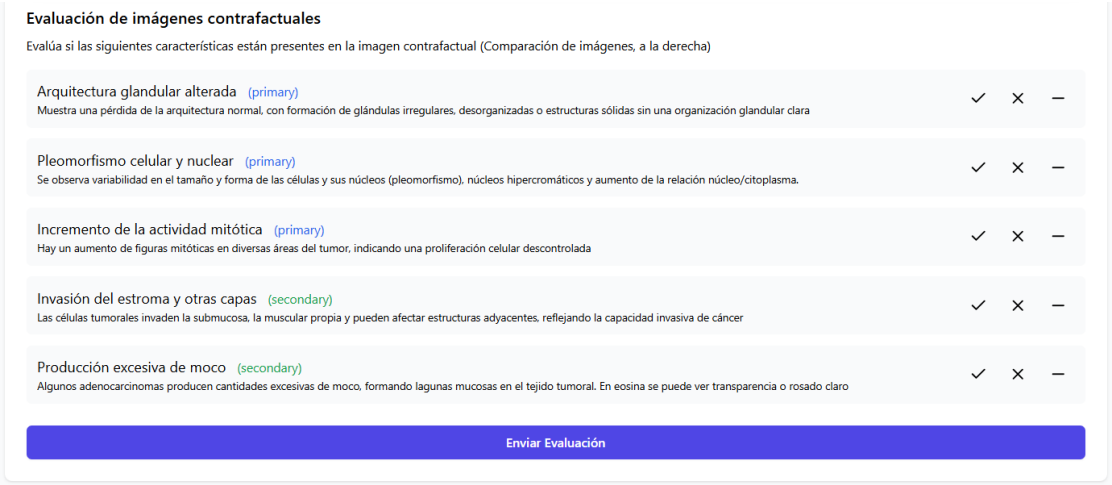


Figura C.6: Pantallazo de la primera evaluación de la sección 2 del software, parte 2.

Figura C.7: Pantallazo de las instrucciones de la sección 3 del software.

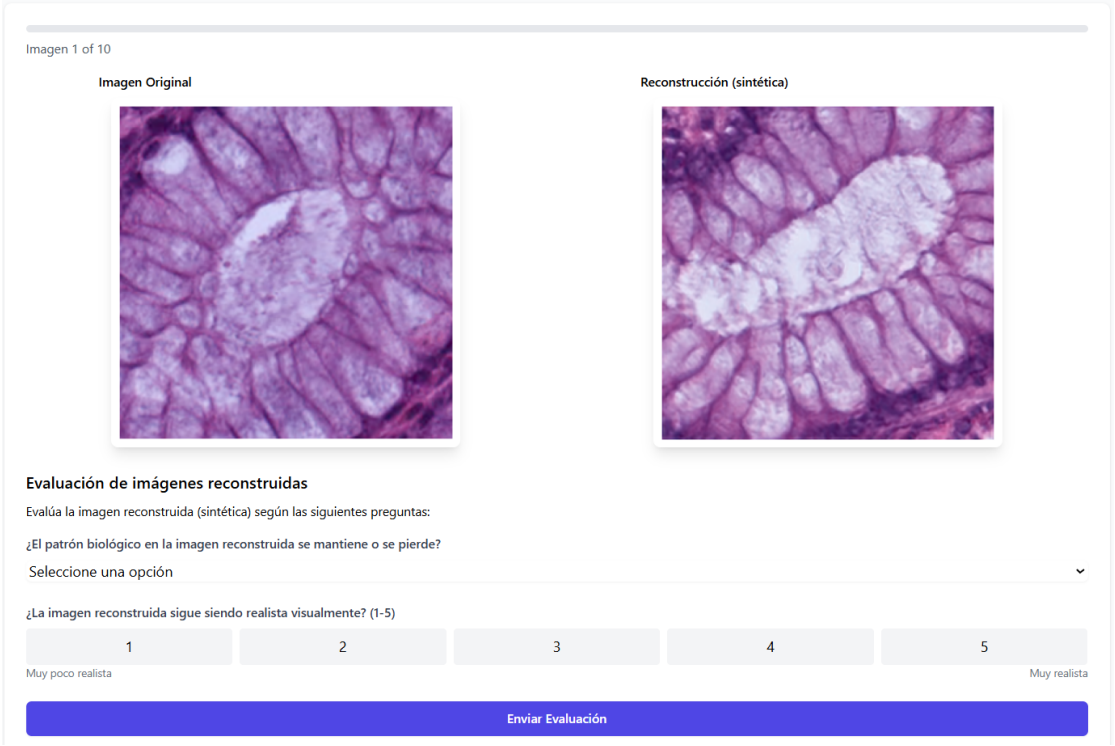


Figura C.8: Pantallazo de la primera evaluación de la sección 3 del software.

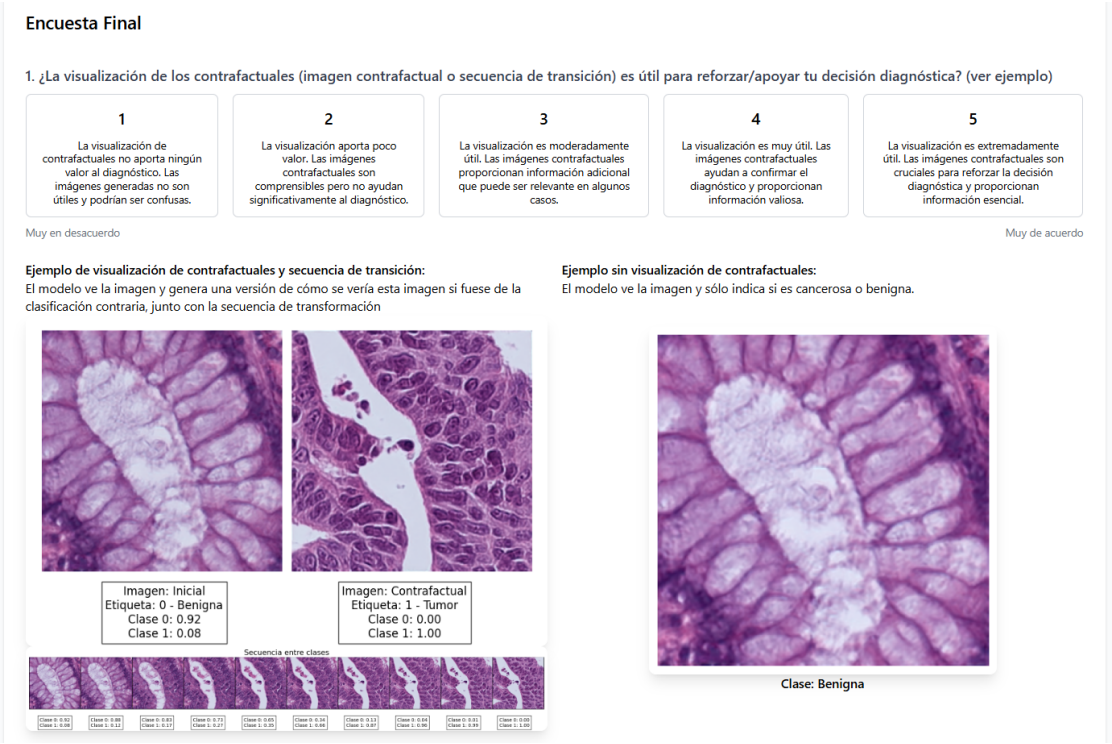


Figura C.9: Pantallazo de la evaluación de la sección final del software, parte 1.

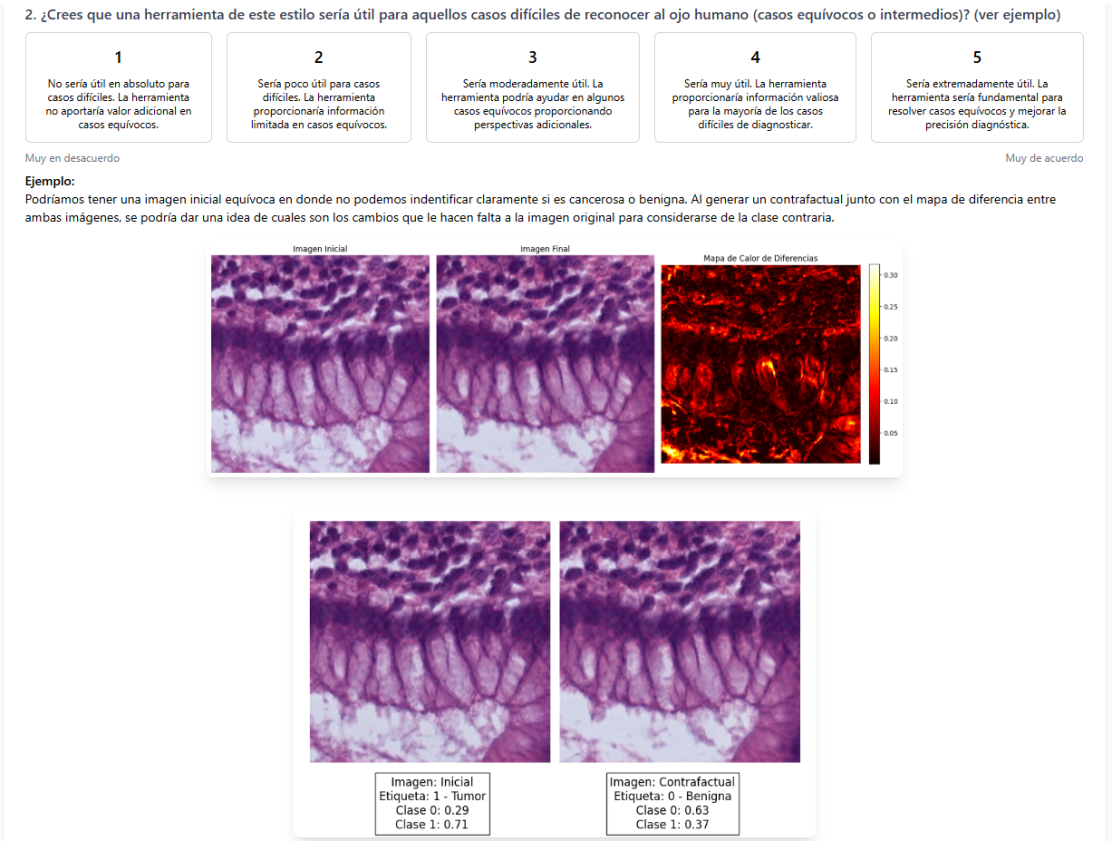


Figura C.10: Pantallazo de la evaluación de la sección final del software, parte 2.

C.2. Resultados Etapa 2: Generación y evaluación de Imágenes Contrafactuales

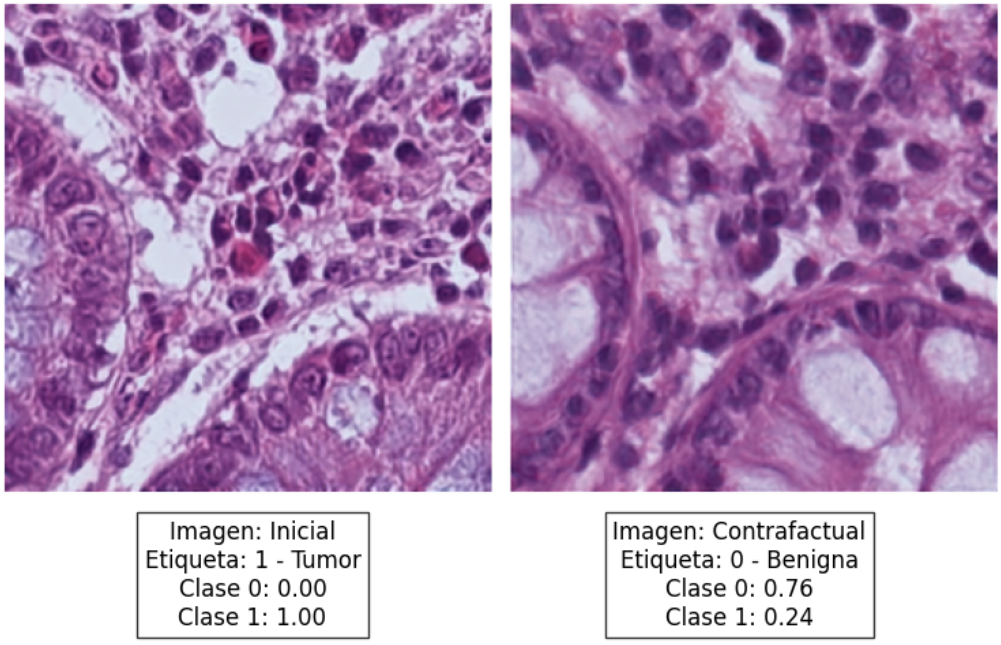


Figura C.11: Imagen 01. Clase Benigno. $S = 68,40\ \%$

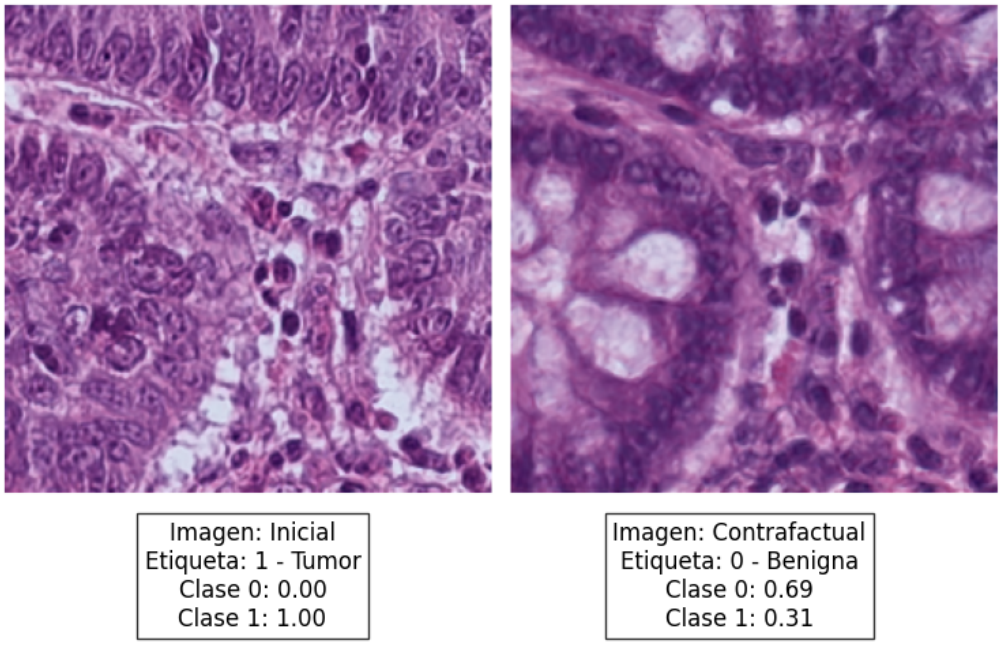


Figura C.12: Imagen 02. Clase Benigno. $S = 67,36\ \%$

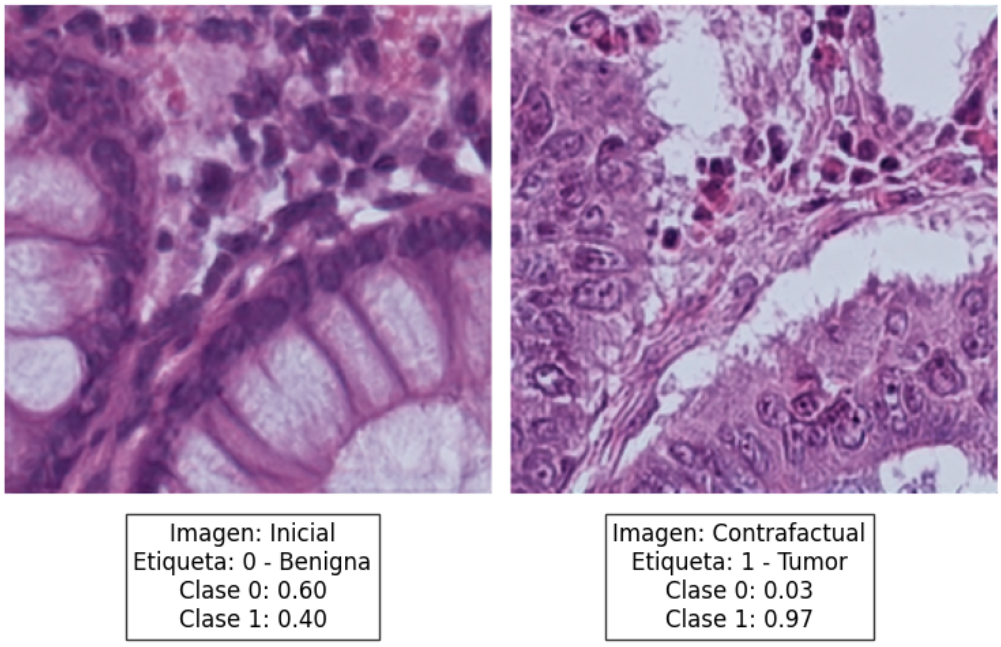


Figura C.13: Imagen 03. Clase Canceroso. $S = 64,93\ \%$

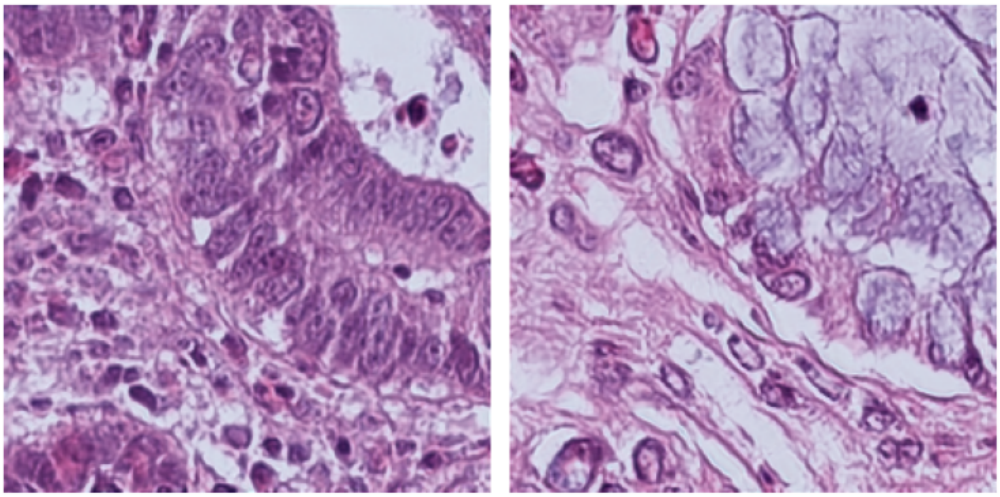


Figura C.14: Imagen 04. Clase Benigno. $S = 77,43\ \%$

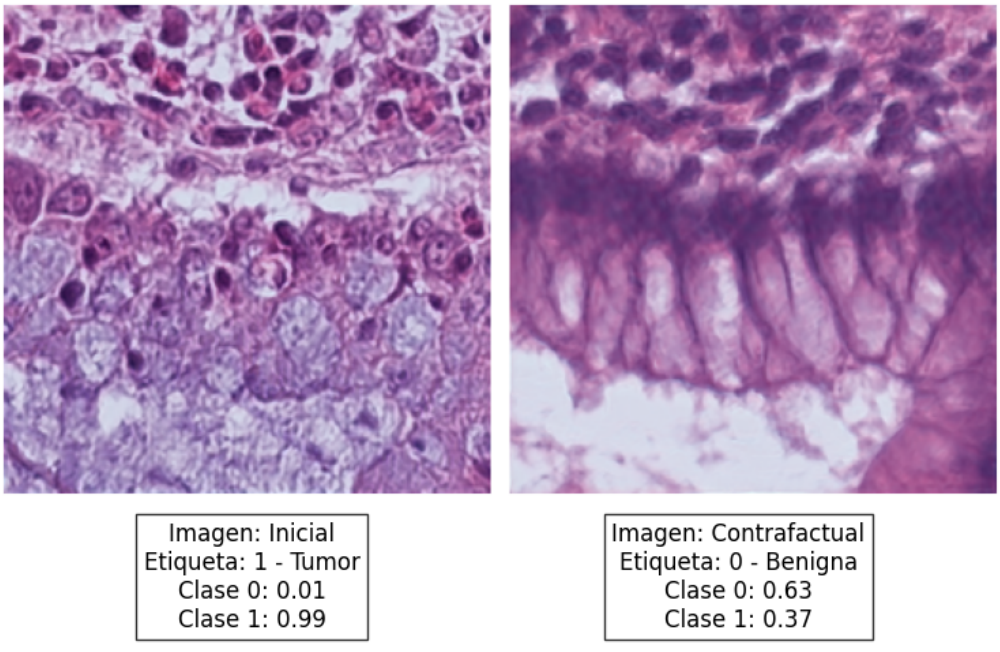


Figura C.15: Imagen 05. Clase Benigno. $S = 51,04\ \%$

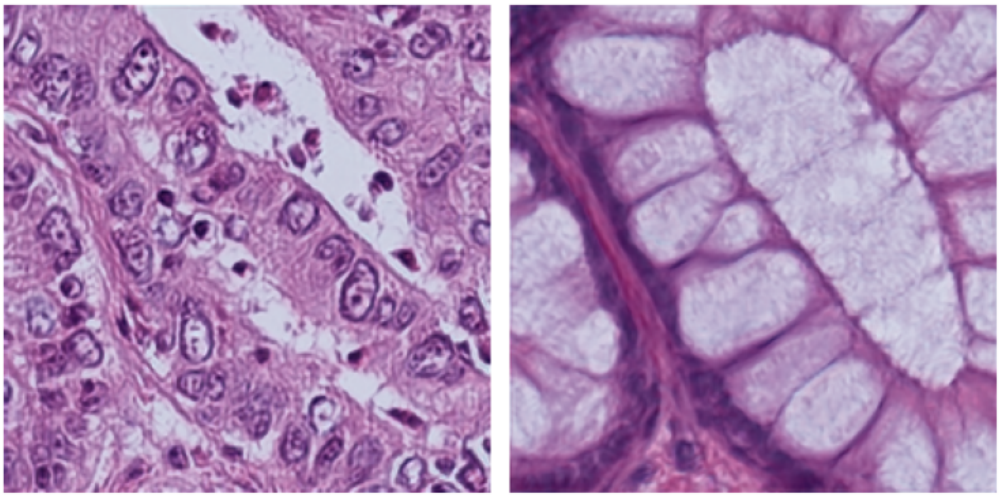


Figura C.16: Imagen 06. Clase Benigno. $S = 29,51\ \%$

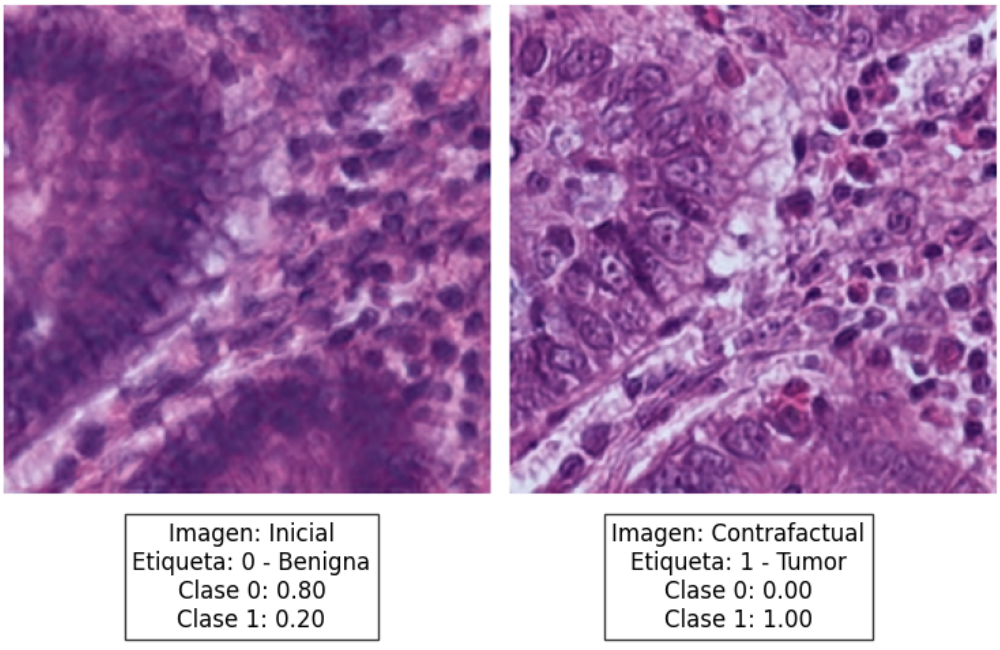


Figura C.17: Imagen 07. Clase Canceroso. $S = 69,44\ \%$

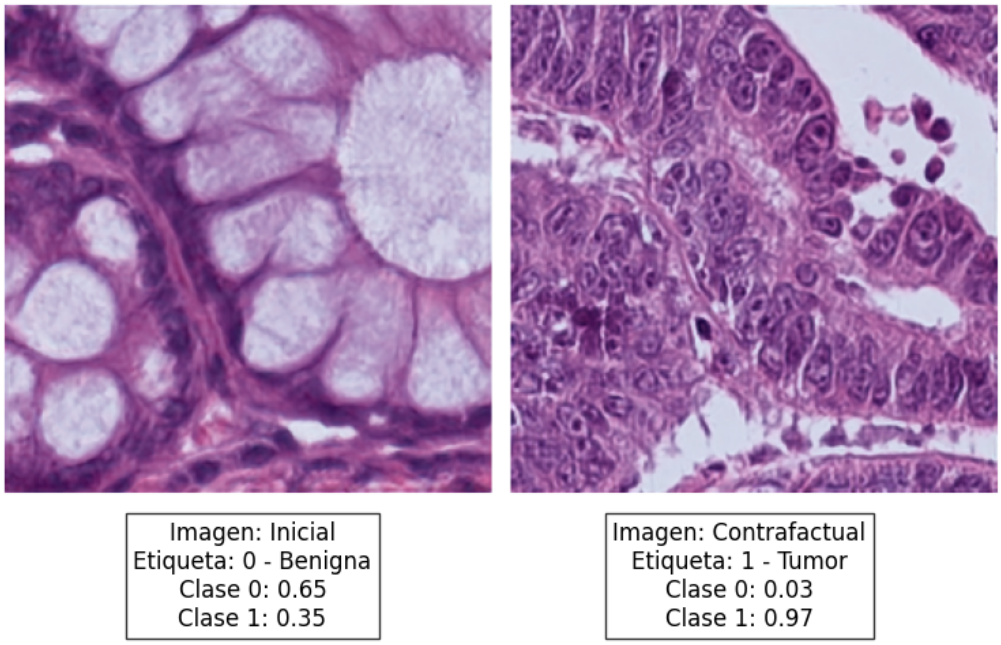


Figura C.18: Imagen 08. Clase Canceroso. $S = 81,75\ \%$

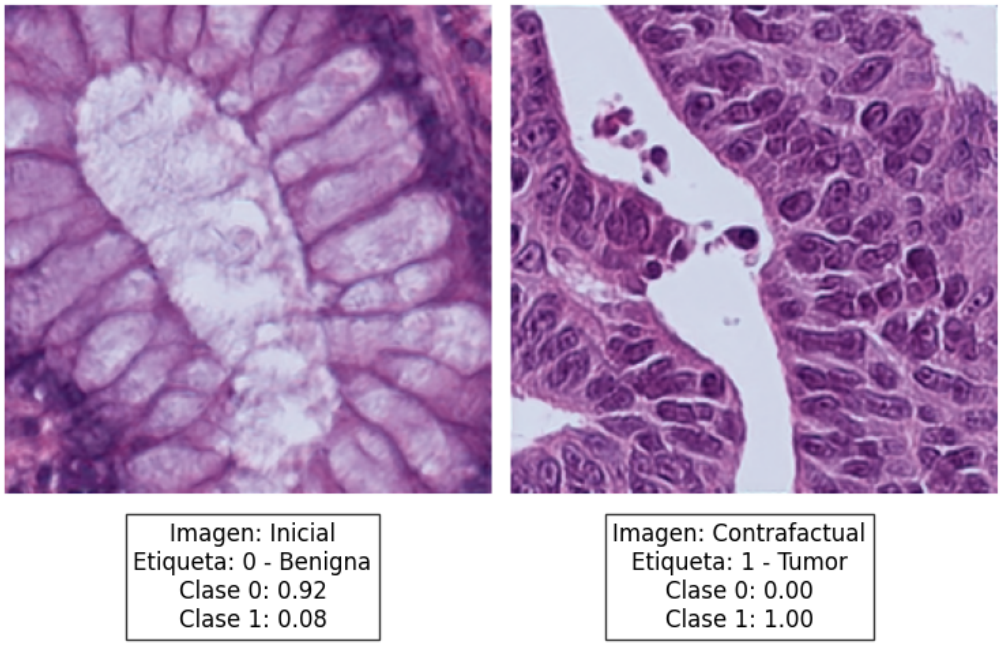


Figura C.19: Imagen 09. Clase Canceroso. $S = 73,02\ \%$

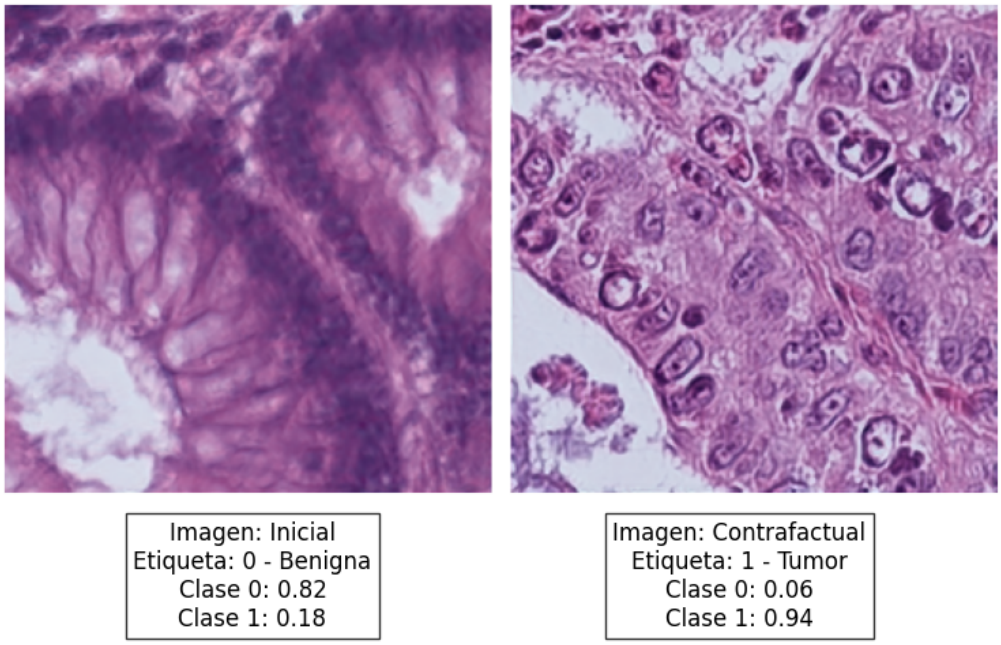


Figura C.20: Imagen 10. Clase Canceroso. $S = 86,11\ \%$

Apéndice D

Histopathology Image Augmentation through StyleGAN2-ADA

Branndon Muñoz[1], Raquel Pezoa[1,2], and Helen Gutierrez[3]

[1] Departamento de Informática, Universidad Técnica Federico Santa María, Valparaíso, Chile
[2] Centro Científico Tecnológico de Valparaíso, Universidad Técnica Federico Santa María, Valparaíso, Chile
[3] Escuela de Tecnología Médica, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile branndon.munoz@sansano.usm.cl, raquel.pezoa@usm.cl, helen.gutierrez@pucv.cl

Abstract. The precise classification of histopathological images is crucial for diagnosing and treating cancer, yet the scarcity of labeled data often limits it. This study investigates the efficacy of using StyleGAN2-ADA data augmentation in histopathological image classification. In this work, we (i) evaluate the capability of StyleGAN2-ADA to generate realistic synthetic histopathological images, (ii) implement a data augmentation strategy using these images, and (iii) compare the performance of a binary classifier trained with and without the proposed augmentation. We trained a StyleGAN2-ADA model in a high-performance computing environment to generate high-quality synthetic images to augment histopathological datasets. We then trained binary classifiers using the augmented datasets for the PCam and IDC datasets and compared their performance with classifiers trained only with original data. Results showed a significant improvement in classifier accuracy, with a 5.9% increase in ROC (AUC) for the PCam dataset at 3% data availability and an 11.3% increase at 20% data availability. For the IDC dataset, the ROC (AUC) improved by 3.4% at 3% data availability and by 2.4% at 20% data availability. Notable enhancements were observed in cancer class metrics, particularly in low-data scenarios, demonstrating the effectiveness of StyleGAN2-ADA in improving classifier robustness and generalization. We conclude that StyleGAN2-ADA is an effective tool for generating high-quality synthetic histopathological images and that the proposed data augmentation strategy substantially enhances classifier performance. Thus, we improved the robustness and generalization of classification models in this critical medical field. Furthermore, it highlights the importance of HPC in accelerating deep learning research applied to complex medical problems.

1 Introduction

Deep learning has revolutionized the analysis of biomedical images, particularly in histopathology, where image assessment is crucial in medical diagnosis [12].

However, histopathological images present unique challenges due to their inherent complexity, exhibiting high variability stemming from differences in tissue samples, staining techniques, and image digitalization conditions [11]. Moreover, labeling or annotating large datasets of histopathological images is costly and time-consuming, presenting a significant barrier to developing robust deep learning-based models in this field [13].

Automatic classification of histopathological images using deep learning-based systems has shown great potential to improve diagnostic accuracy and reduce pathologists' workloads. However, developing robust and generalizable models remains challenging due to several factors. The scarcity of labeled data is particularly critical for rare or underrepresented pathologies. Furthermore, class imbalance in the available datasets can lead to biases in the trained models, compromising their performance and reliability [18]. To address these challenges, generating synthetic data using techniques such as Generative Adversarial Networks (GANs), first proposed by Goodfellow et al. in 2014 [4], has emerged as a promising solution. GANs are unsupervised machine learning models capable of generating synthetic images for use in training, offering the potential to create high-quality synthetic histopathological images and thereby expand the datasets available for training classification models [21]. However, the application of GANs in the context of histopathology is not without its own set of challenges. These include instability during training, a predisposition to mode collapse, and the need to ensure that the generated images accurately represent real samples [2].

Despite the mentioned difficulties, diverse works have demonstrated the successful use of GANs in digital pathology with promising results. Alajaji et al. [1] presented a comprehensive overview of the applications and limitations of GANs in the histopathology field. Xue et al. [20] showed how selective synthetic augmentation can improve the classification of histopathological images. In this context, StyleGAN2-ADA (Adaptive Discriminator Augmentation) [8] emerges as an advanced GAN architecture that has the potential to overcome some of the aforementioned limitations. StyleGAN2-ADA incorporates adaptive augmentation techniques that improve training stability and the quality of generated images, making it particularly suitable for generating synthetic histopathological images.

This work addresses the challenge of data augmentation for classifying histopathological images. We propose a method capable of generating high-quality synthetic histopathological images using StyleGAN2-ADA and incorporating them into a dynamic training process for image classification. The main contributions of this work include: (i) an evaluation of StyleGAN2-ADA's efficacy in generating high-quality and representative synthetic histopathological images, (ii) a novel method for data augmentation in digital pathology, utilizing images generated by StyleGAN2-ADA to enrich and balance training datasets, and (iii) a comprehensive comparison of the performance of a binary histopathological image classifier trained with and without the proposed data augmentation technique, quantifying the impact on model accuracy and robustness.

2 Related Work and Background

2.1 Histopathological image classification

The classification of histopathological images has experienced significant advances in recent years, driven by progress in deep learning. However, this field faces unique challenges that have motivated the exploration of advanced data augmentation and image synthesis techniques. Traditional approaches in histopathological analysis have employed various supervised learning methods. Chankong et al. [3] proposed the automatic segmentation and classification of cells using fuzzy C-means clustering for cervical cancer, while Guo et al. [15] designed hand-crafted nucleus-based features for the classification of digitized epithelium.

The application of convolutional neural networks (CNNs) at the patch level, as demonstrated by Hou et al. [6], has been a common approach to handle the high resolutions of whole slide images (WSI). However, the accuracy in classifying these patches is crucial to achieving performance comparable to human pathologists. More recently, Tellez et al. [17] addressed the challenge of variability in the appearance of histopathological images due to differences in sample preparation and digitization, proposing a normalization method based on generative adversarial networks (GANs) to improve generalization across different centers and scanners. Despite these advances, histopathological image classification still faces significant challenges, including the scarcity of labeled data, class imbalance, and model interpretability [5].

2.2 Synthetic data augmentation

Synthetic data augmentation has emerged as a powerful technique to address the challenges of data scarcity and class imbalance in histopathological image classification. This technique goes beyond traditional augmentation methods by generating new synthetic samples that expand the diversity of the training dataset. Mahmood et al. [14] demonstrated the efficacy of synthetic data augmentation in histopathology using conditional GANs to generate colorectal tissue images, improving classification accuracy and model interpretability. Meanwhile, Xue et al. [20] proposed a selective synthetic augmentation approach using HistoGAN, a GAN specifically designed for histopathological images, significantly improving breast cancer classification accuracy. However, the generation of high-quality synthetic data for augmentation in histopathology has not been fully realized. Typically, synthetic samples are blindly added to the original data without using architectures capable of capturing fine details, preserving structural coherence, or ensuring diversity, representativeness, and control of augmentation in the generated samples.

The StyleGAN architecture, proposed by Karras et al. [9], and its improved version StyleGAN2 [10], have demonstrated exceptional capabilities in generating high-resolution and detailed images. StyleGAN2 introduced significant improvements, including a revised generator design, weight normalization, and a new path length regularizer for more stable training. More recently, Karras et

al. [8] introduced StyleGAN2-ADA (Adaptive Discriminator Augmentation), an extension that addresses the challenge of training GANs with limited datasets. ADA dynamically adjusts the intensity of augmentation during training, enabling the generation of high-quality images even with small datasets, a particularly valuable feature in histopathology.

Our work builds on these previous advances by applying StyleGAN2-ADA specifically to the task of synthesizing histopathological images for data augmentation. By leveraging StyleGAN2-ADA's ability to generate diverse and high-quality images, we aim to address persistent challenges in histopathological image classification, including labeled data scarcity, class imbalance, and variability in sample preparation. Additionally, we implement a rigorous validation process that includes both quantitative metrics and qualitative evaluation by clinical experts, ensuring that our synthetic data are not only visually convincing but also clinically relevant and useful for improving the performance of classification models.

3 Methods

Our approach addresses the challenge of limited labeled data in histopathological image classification by leveraging StyleGAN2-ADA for synthetic data generation and incorporating it into a dynamic training process for a ResNet34-based classifier. This method aims to improve classification performance by augmenting the training dataset with high-quality synthetic images. The methodology comprises two main components: synthetic data generation and classifier training with dynamic augmentation.

Synthetic Data Generation using StyleGAN2-ADA. StyleGAN2-ADA serves as the backbone of our synthetic data generation process. We chose this architecture for its ability to produce high-quality images even when trained on limited datasets, a common scenario in histopathological imaging. StyleGAN2-ADA builds upon the StyleGAN2 architecture, incorporating adaptive discriminator augmentation to prevent overfitting when training on small datasets. The model consists of a mapping network that transforms input latent codes into an intermediate latent space, and a synthesis network that generates images based on these intermediate latent codes.
The adaptive discriminator augmentation mechanism was crucial in stabilizing the training process and improving the quality of generated images. We monitored the Fréchet Inception Distance (FID) score throughout the training to assess the quality of generated images and selected the model checkpoint with the lowest FID score for our synthetic data generation.

We trained the StyleGAN2-ADA model on two histopathological datasets using the configuration shown in Table 1.

Once the StyleGAN2-ADA models were trained, we trained a classifier based on a pre-trained ResNet34 architecture. We modified the last layers to suit our

Resolution	Kimg	Learning Rate	R1 gamma	Augments
128×128	25,000	0.0025 Adam	0.0256	bgc

Table 1: Training configuration of StyleGAN2-ADA models used in this work. The augment tipe bgc refers to blit, geom and color augments [8].

binary classification task of histopathological tissues. We added dynamic augmentation that combines real and synthetic images to form the training batch. This approach ensures that the model is continuously exposed to new synthetic data throughout the training process, potentially improving its generalization capabilities. The augmentation is performed according to a percentage r of images per class, that is, with $r = 0.5$ we should generate 50% of the amount of images for each class.

To perform a comparison against a baseline, we trained a classifier with an augmentation strategy different from ours. This strategy uses class-weights [16] and traditional data augmentation, including color jitter and random horizontal flip. Class-weights are coefficients applied to each class during model training to balance the influence of each class in the loss function. The goal is to penalize errors in minority classes more and penalize errors in majority classes less. Table 2 shows the base configuration for each classifier without data augmentation.

Loss function	Optimizer	Learning Rate	Batch-size	Epochs
Binary Cross-Entropy + class-weights	SGD	$[1e{-}6, 0.1]$	32	30

Table 2: Training configuration of baseline classifier pretrained on ResNet34 used in this work.

We used the following metrics for evaluation; Accuracy, Precision, Recall, F1-score and Area Under the Receiver Operating Characteristic curve (AUC-ROC). Through this methodology, our goal is to demonstrate the potential of combining StyleGAN2-ADA for synthetic data generation with dynamic data augmentation to improve the performance of histopathological image classification, particularly in scenarios with limited labeled data.

4 Results

4.1 Datasets

For this study, two histopathological image datasets were used, each presenting unique challenges in histopathological image classification. The Invasive Ductal Carcinoma (IDC) dataset [7] is derived from 162 whole slide images of breast cancer specimens scanned at 40x magnification. Instead of using the entire slide images, 277,524 patches of 50x50 pixels were extracted, with 198,738 (71.6%) negative for IDC and 78,786 (28.4%) positive for IDC. Each patch was labeled binary, where 0 indicates the absence of IDC and 1 indicates the presence of IDC. The particular challenge with this dataset is the class imbalance, as it presents approximately 2.5 times more negative samples (absence of IDC) than positive samples (presence of IDC). The dataset was randomly split into training, validation, and test sets, at a ratio of 7:1:2, respectively, maintaining the same imbalance in each set. All comparisons and evaluations were carried out considering the test set.

To provide an additional validation dataset, we also experimented with the public PatchCamelyon (PCam) dataset [19]. PCam consists of 327,680 color patches of 96x96 pixels extracted from histopathological scans of lymph node sections. Each patch is annotated with a binary label indicating the presence or absence of metastatic tissue. Unlike the IDC dataset, the PCam dataset is balanced, with an equal representation of the negative class (absence of metastatic tissue) and the positive class (presence of metastatic tissue). The dataset is divided into training, validation, and test sets at a ratio of 75%:12.5%:12.5%, respectively, maintaining this balance across all sets.

Testing with a balanced dataset like PCam and an imbalanced dataset like IDC is essential to evaluate the robustness and generalizability of our classification models. The balanced nature of PCam allows us to observe model performance under ideal conditions, while the imbalanced IDC dataset presents real-world challenges, such as class imbalance, which can significantly affect model predictions. Intentionally altering the balance of PCam would compromise its role as a benchmark dataset, making it unsuitable for this comparative analysis. Maintaining the natural balance in PCam is crucial for understanding how models perform when trained on datasets that do not inherently favor one class over the other.

4.2 Experiments

To simulate the situation where we only have a limited amount of training data available, 3% and 20% of the images from the training set of each dataset were randomly selected to train 4 StyleGAN2-ADA models for augmentation. The configuration and the FID metric results obtained can be seen in Table 3

Additionally, two classifiers were trained for each dataset using 3% and 20% of the images from their respective training sets (4 classifiers with data augmentation using StyleGAN2-ADA) and two classifiers with the baseline configuration

Dataset	Resolution	# images	FID	Training time
PCam (subset)	128×128	52429 (20%)	3.200	3d 09h 06m
PCam (subset)	128×128	7864 (3%)	3.738	1d 02h 24m
IDC (subset)	64×64	37789 (20%)	2.608	1d 14h 00m
IDC (subset)	64×64	5668 (3%)	2.603	0d 15h 09m

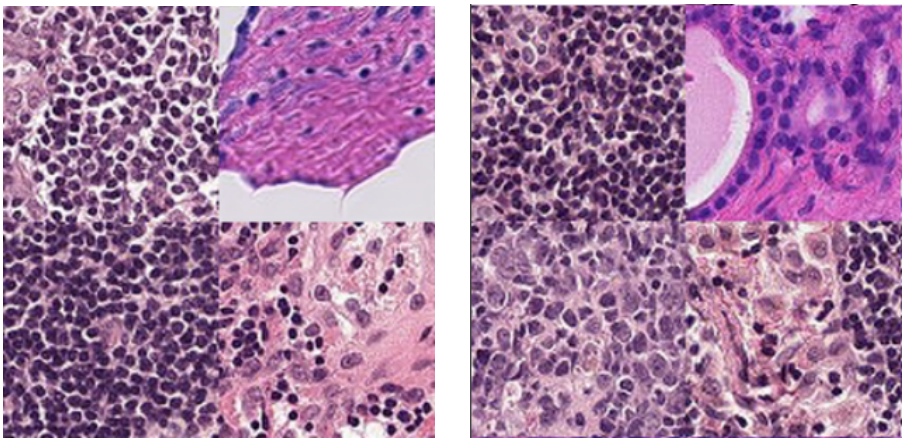
Table 3: Configuration and training results of StyleGAN2-ADA with 2 different datasets; PCam and IDC. The training was executed on a single V100 GPU from NLHPC (National Laboratory for High Performance Computing).

for each dataset on the same sets (3% and 20% of the training set images). In all classifiers with data augmentation generated with StyleGAN2-ADA, the same configuration as the baseline model was used but without class-weights or traditional data augmentation, which includes color jitter and random horizontal flip. The data augmentation rate was set to $r = 0.3$ for all classifiers with augmentation. Additionally, since IDC is a highly imbalanced dataset, unlike PCam, in addition to applying data augmentation with $r = 0.3$, we generated an amount of minority class data equal to the amount of majority class data.

4.3 Result Analysis

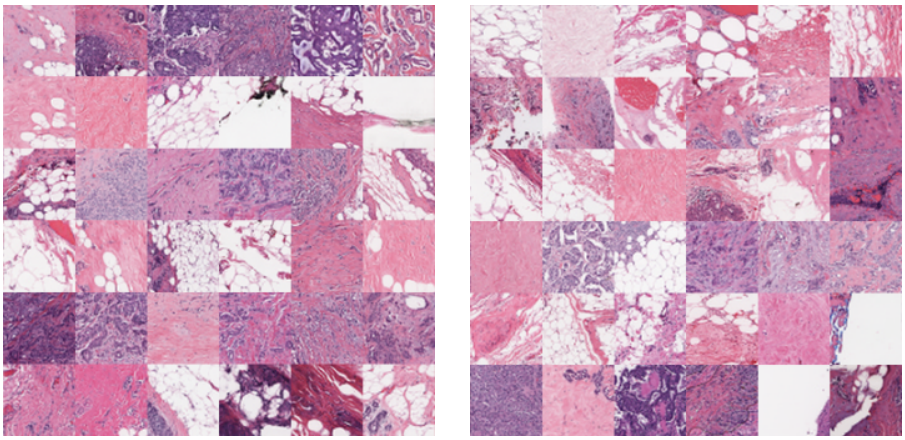
Qualitative evaluation. Figures 1 and 2 compare synthetic images generated with StyleGAN2-ADA and real images for the PCam and IDC datasets. In both datasets, despite perceiving an instant visual similarity, it is not correct to determine by simple human observation whether these serve for data augmentation or not without domain knowledge and/or being backed by a quantitative evaluation.

To assess the quality of the images we generated and validate the effectiveness of the synthetic augmentation method, we invited a clinical specialist to perform an expert qualitative evaluation from a histopathological perspective on both datasets, namely, PCam and IDC. The real images and those generated by the StyleGAN2-ADA model were compared. For this, the specialist was provided with two grids of 32×32 generated images and two grids of real images from both datasets for visual inspection. According to their criteria, the generated images are coherent and quite similar to the real ones. As part of our future work, we will elaborate a survey to rank the images on a scale from 1 (worst) to 5 (best) and ask a panel of specialists for their opinions. The generators used for generating these grids obtained an FID score of 3.20 and 2.608 for PCam and IDC, respectively.



(a) Synthetic images (Fakes) (b) Real Images

Fig. 1: Example of real and generated images from the PCam dataset. The obtained FID score is 3.2.



(a) Synthetic images (Fakes) (b) Real Images

Fig. 2: Example of real and generated images from the IDC dataset. The obtained FID score is 2.603.

Quantitative evaluation. As we already mentioned, we evaluated the performance of StyleGAN2-ADA data augmentation on two datasets: PCam (PatchCamelyon) and IDC (Invasive Ductal Carcinoma). Our experiments focused particularly on binary classification tasks (cancer vs. non-cancer) at two different

percentages of data availability: 3% and 20%. We compared the baseline model performance against models trained with StyleGAN2-ADA augmentation, using various metrics including ROC (AUC), accuracy, recall, and F1-score for both cancer and non-cancer classes.

PCam Dataset Results. Table 4 presents the classification results for the PCam dataset. At 3% data availability, the baseline model achieved a ROC (AUC) of 0.813, while the StyleGAN2-ADA augmented model improved this to 0.861, representing a 5.9% increase. The augmented model also showed improvements across all other metrics for both classes, with notable enhancements in cancer class accuracy (from 0.86 to 0.89) and recall (from 0.70 to 0.81). At 20% data availability, the baseline ROC (AUC) was 0.797, which increased to 0.887 with StyleGAN2-ADA augmentation, marking an 11.3% improvement. The augmented model maintained or slightly improved most metrics, with the most significant gains in cancer class accuracy (from 0.82 to 0.84) and F1-score (from 0.78 to 0.80).

PCam	Model	ROC AUC	Class: Non-cancer			Class: Cancer		
			Accuracy	Recall	F1-score	Accuracy	Recall	F1-Score
3%	Baseline	0.813	0.75	0.88	0.81	0.86	0.70	0.77
	+ StyleGAN2-ADA augmentation, r = 0.3 + balance	**0.861**	**0.86**	**0.89**	**0.87**	**0.89**	**0.81**	**0.85**
20%	Baseline	0.797	0.77	0.83	0.80	0.82	0.75	0.78
	+ StyleGAN2-ADA augmentation, r = 0.3 + balance	**0.887**	**0.78**	**0.85**	**0.81**	**0.84**	**0.76**	**0.80**

Table 4: Classification results with data augmentation using StyleGAN2-ADA at different percentages of PCam dataset, separated by class (cancer/non-cancer).

IDC Dataset Results. Table 5 shows the classification results for the IDC dataset. At 3% data availability, the baseline model achieved a ROC (AUC) of 0.856, while the StyleGAN2-ADA augmented model improved this to 0.885, a 3.4% increase. Notably, the augmented model substantially improved the cancer class metrics, increasing accuracy from 0.75 to 0.89, recall from 0.63 to 0.85, and F1-score from 0.69 to 0.87. At 20% data availability, the baseline ROC (AUC) was 0.862, which increased to 0.883 with StyleGAN2-ADA augmentation, a 2.4% improvement. However, a detailed analysis reveals that while the cancer class metrics significantly improved (accuracy from 0.81 to 0.88, recall from 0.56 to

0.84, and F1-score from 0.66 to 0.86), the metrics for the non-cancer class decreased. This decrease in the non-cancer class metrics should not be viewed as a decline in model performance. Initially, the baseline model was likely biased towards the majority class (non-cancer), leading to seemingly good metrics that did not fully represent its true performance. This trade-off between the metrics of different classes is expected and often desirable when addressing class imbalance. The overall increase in the ROC (AUC) further confirms that the improved model is more discriminative and effective for the classification task. Additionally, in a medical context, this improved detection of positive cases (cancer) is generally more valuable than maintaining a higher precision for negative cases (non-cancer).

These results demonstrate that StyleGAN2-ADA data augmentation consistently improves model performance across both datasets, with particularly pronounced benefits for the cancer class in the IDC dataset. The augmentation technique proves especially effective in low-data scenarios (3% availability), but also shows improvements when more data is available (20%).

IDC	Model	ROC AUC	Class: Non-cancer			Class: Cancer		
			Accuracy	Recall	F1-score	Accuracy	Recall	F1-Score
3%	Baseline	0.856	**0.89**	**0.93**	**0.91**	0.75	0.63	0.69
	+ StyleGAN2-ADA augmentation, r = 0.3 + balance	**0.885**	0.85	0.89	0.87	**0.89**	**0.85**	**0.87**
20%	Baseline	0.862	**0.87**	**0.96**	**0.91**	0.81	0.56	0.66
	+ StyleGAN2-ADA augmentation, r = 0.3 + balance	**0.883**	0.84	0.90	0.87	**0.88**	**0.84**	**0.86**

Table 5: Classification results with data augmentation using StyleGAN2-ADA at different percentages of IDC dataset, separated by class (cancer/non-cancer).

4.4 Discussion.

The results of this study clearly demonstrate the efficacy of StyleGAN2-ADA in generating high-quality synthetic histopathological images and the substantial improvements it brings to image classification performance. The data augmentation strategy proposed in this research shows significant benefits, particularly in scenarios with limited data availability, which is a common challenge in the field of histopathology. In the IDC dataset at 20% data availability, while the metrics for the cancer class improved significantly, the metrics for the non-cancer class decreased. However, this reduction does not indicate worse performance. On the

contrary, it reflects that the model is now less biased toward the majority class (non-cancer), making it more discriminative and effective for the classification task. The improvements in metrics such as accuracy, recall, and F1-score, especially for the cancer class, highlight the potential of this approach to enhance the robustness and generalization of classification models. This is particularly important in medical applications where accurate and reliable classification is critical. Looking forward, we aim to further exploit StyleGAN2-ADA to generate data in a more specific manner by having control over the attributes of the synthetic images. Additionally, we plan to adopt a selective approach to filter out synthetic data that falls outside the target distribution. By implementing these strategies, we believe we can further enhance the results and address the class imbalance observed in our current study. HPC resources were instrumental in accelerating the training and generation processes, underscoring the importance of computational power in advancing deep learning research in complex medical domains. In conclusion, this study demonstrates that StyleGAN2-ADA is a powerful tool for addressing the data scarcity challenge in histopathology and significantly improving the performance of classification models. The findings have important implications for developing robust and generalizable models in medical image analysis, ultimately contributing to better diagnostic and treatment outcomes in cancer care.

4.5 Implementation details

All models were created using PyTorch. The generative models were trained on a single NVIDIA Tesla V100 GPU, and the classifiers were trained on a T4 GPU from Google Colab Pro+. The specific implementation details are the same as those used by StyleGAN2-ADA and can be found in the following repository.
The total training time is significantly influenced by factors such as resolution, the number of GPUs, the dataset, the desired output quality, and the selected hyperparameters. The Table 6 outlines the expected wallclock times to reach various stages in the training process, measured in thousands of real images shown to the discriminator ("kimg"). Typically, it takes around 25000 kimg or more to achieve full convergence, though results are often quite satisfactory at around 5000 kimg.

The utilization of high-end GPUs like the NVIDIA Tesla V100 is essential for projects of this nature, as training models with high-resolution images becomes computationally prohibitive with standard GPUs. Access to such advanced hardware is often limited on more affordable platforms, while those offering extended usage are prohibitively expensive. As demonstrated, even with these top-tier GPUs, training on higher resolution images, such as 1024 pixels, can still be very time-consuming, underscoring the need for advanced computational resources in these projects.

The source code and datasets used in this study will be made available in this public Github repository to facilitate the reproducibility of the results.

Resolution	GPUs	1000 kimg	25000 kimg	sec/kimg	GPU mem	CPU mem
128×128	1	4h 05m	4d 06h	12.8–13.7	7.2 GB	3.9 GB
	2	2h 06m	2d 04h	6.5–6.8	7.4 GB	7.9 GB
	4	1h 20m	1d 09h	4.1–4.6	4.2 GB	16.3 GB
256×256	1	6h 36m	6d 21h	21.6–24.2	5.0 GB	4.5 GB
	2	3h 27m	3d 14h	11.2–11.8	5.2 GB	9.0 GB
	4	1h 45m	1d 20h	5.6–5.9	5.2 GB	17.8 GB
512×512	1	21h 03m	21d 22h	72.5–74.9	7.6 GB	5.0 GB
	2	10h 59m	11d 10h	37.7–40.0	7.8 GB	9.8 GB
	4	5h 29m	5d 17h	18.7–19.1	7.9 GB	17.7 GB
1024×1024	1	1d 20h	46d 03h	154.3–161.6	8.1 GB	5.3 GB
	2	23h 09m	24d 02h	80.6–86.2	8.6 GB	11.9 GB
	4	11h 36m	12d 02h	40.1–40.8	8.4 GB	21.9 GB

Table 6: StyleGAN2-ADA [10] training configurations and resource usage for different resolutions using various numbers of GPUs.

Acknowledgements. The authors acknowledge the financial support from ANID PIA/APOYO AFB230003, ANID FONDECYT Postdoc Project and 3190740 and Universidad Técnica Federico Santa María for Beca Financiera Magíster. Special thanks are extended to the National Laboratory for High Performance Computing (NLHPC) for providing free access to their computing cluster, including storage and GPUs, without which this project would not have been possible.

References

1. Alajaji, S.A., Khoury, Z.H., Elgharib, M., Saeed, M., Ahmed, A.R., Khan, M.B., Tavares, T., Jessri, M., Puche, A.C., Hoorfar, H., Stojanov, I., Sciubba, J.J., Sultan, A.S.: Generative adversarial networks in digital histopathology: Current applications, limitations, ethical considerations, and future directions. Modern Pathology **37**(1), 100369 (2024). https://doi.org/https://doi.org/10.1016/j.modpat.2023.100369, https://www.sciencedirect.com/science/article/pii/S0893395223002740
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein gan (2017), https://arxiv.org/abs/1701.07875
3. Chankong, T., Theera-Umpon, N., Auephanwiriyakul, S.: Automatic cervical cell segmentation and classification in pap smears. Computer methods and programs in biomedicine **113 2**, 539–56 (2014), https://api.semanticscholar.org/CorpusID:478558
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

5. Graziani, M., Andrearczyk, V., Müller, H.: Regression concept vectors for bidirectional explanations in histopathology (2019), https://arxiv.org/abs/1904.04520
6. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification (2016), https://arxiv.org/abs/1504.07947
7. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of Pathology Informatics **7**(1), 29 (2016). https://doi.org/https://doi.org/10.4103/2153-3539.186902, https://www.sciencedirect.com/science/article/pii/S2153353922005478
8. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in neural information processing systems **33**, 12104–12114 (2020)
9. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019), https://arxiv.org/abs/1812.04948
10. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan (2020), https://arxiv.org/abs/1912.04958
11. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. Computational and Structural Biotechnology Journal **16**, 34–42 (2018). https://doi.org/https://doi.org/10.1016/j.csbj.2018.01.001, https://www.sciencedirect.com/science/article/pii/S2001037017300867
12. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. CoRR **abs/1702.05747** (2017), http://arxiv.org/abs/1702.05747
13. Madabhushi, A., Lee, G.: Image analysis and machine learning in digital pathology: Challenges and opportunities. Medical Image Analysis **33**, 170–175 (2016). https://doi.org/https://doi.org/10.1016/j.media.2016.06.037, https://www.sciencedirect.com/science/article/pii/S1361841516301141, 20th anniversary of the Medical Image Analysis journal (MedIA)
14. Mahmood, F., Borders, D., Chen, R., McKay, G.N., Salimian, K.J., Baras, A., Durr, N.J.: Deep adversarial training for multi-organ nuclei segmentation in histopathology images (2018), https://arxiv.org/abs/1810.00236
15. Sohail, A., Khan, A., Nisar, H., Tabassum, S., Zameer, A.: Mitotic nuclei analysis in breast cancer histopathology images using deep ensemble classifier. Medical Image Analysis **72**, 102121 (2021). https://doi.org/https://doi.org/10.1016/j.media.2021.102121, https://www.sciencedirect.com/science/article/pii/S1361841521001675
16. Sun, Y., Wong, A.K.C., Kamel, M.S.: Classification of imbalanced data: a review. Int. J. Pattern Recognit. Artif. Intell. **23**, 687–719 (2009), https://api.semanticscholar.org/CorpusID:27118324
17. Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J.: Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Medical Image Analysis **58**, 101544 (2019). https://doi.org/https://doi.org/10.1016/j.media.2019.101544, https://www.sciencedirect.com/science/article/pii/S1361841519300799

18. Tizhoosh, H.R., Pantanowitz, L.: Artificial intelligence and digital pathology: Challenges and opportunities. Journal of Pathology Informatics **9**(1), 38 (2018). https://doi.org/https://doi.org/10.4103/jpi.jpi_53_18, https://www.sciencedirect.com/science/article/pii/S2153353922003510
19. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology (Jun 2018)
20. Xue, Y., Ye, J., Zhou, Q., Long, L.R., Antani, S., Xue, Z., Cornwell, C., Zaino, R., Cheng, K.C., Huang, X.: Selective synthetic augmentation with histogan for improved histopathology image classification. Medical Image Analysis **67**, 101816 (2021). https://doi.org/https://doi.org/10.1016/j.media.2020.101816, https://www.sciencedirect.com/science/article/pii/S1361841520301808
21. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: A review. Medical Image Analysis **58**, 101552 (2019). https://doi.org/https://doi.org/10.1016/j.media.2019.101552, https://www.sciencedirect.com/science/article/pii/S1361841518308430